TESIS – TE142599

PREDIKSI KUNJUNGAN HALAMAN WEBSITE DENGAN N-GRAM MODEL

ELOK SRI WAHYUNI
2213206714

DOSEN PEMBIMBING :

Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.

PROGRAM STUDI MAGISTER
BIDANG KEAHLIAN TELEMATIKA
KONSENTRASI CIO
JURUSAN TEKNIK ELEKTRO – FAKULTAS TEKNOLOGI INDUSTRI
INSTITUT TEKNOLOGI SEPULUH NOPEMBER
SURABAYA
2015

THESIS – TE142599

WEB ACCESS PREDICTION USING N-GRAM MODEL

ELOK SRI WAHYUNI
2213206714

Supervisor :
DR. Ir. Yoyon K. Suprapto, M.Sc

MAGISTER PROGRAM
TELEMATICS ENGINEERING
CONCENTRATION CHIEF INFORMATION OFFICER (CIO)
ELECTRICAL ENGINEERING
DEPARTEMENT FACULTY INDUSTRIAL TECHNOLOGY
TENTH of NOPEMEBER INSTITUTE of TECHNOLOGY
SURABAYA
2015

Tesis disusun untuk memenuhi salah satu syarat memperoleh gelar
Magister Teknik (MT)
di
Institut Teknologi Sepuluh Nopember

Oleh :
Elok Sri Wahyuni
2213 206 714

Tanggal Ujian : 15 Januari 2015
Periode Wisuda : Maret 2015

Disetujui Oleh:

1. Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc. (Pembimbing)
   NIP.195409251978031001

2. Dr. Eko Mulyanto Yuniarno, ST.,MT. (Penguji)
   NIP.196806011995121009

3. Dr. I Ketut Eddy Purnama, ST.,MT. (Penguji)
   NIP.196907031995121001

4. Dr. Supeno Mardi Susiki Nugroho, ST.,MT. (Penguji)
   NIP.197003131995121001

Direktur Program Pascasarjana

Prof.Dr.Ir. Adi Soeprijanto, MT
NIP. 19640405 199002 1 001

PREDIKSI KUNJUNGAN HALAMAN WEBSITE DENGAN N-GRAM MODEL

Nama Mahasiswa	:	Elok Sri Wahyuni
NRP	:	2213206714
Pembimbing	:	Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.

ABSTRAK

Prediksi dan pemodelan pola kunjungan pengguna website dapat diukur kinerjanya dengan beberapa parameter. Parameter pengukuran yang sering digunakan yaitu kompleksitas model, kemampuan model dalam membuat prediksi (*aplicability*) dan akurasi prediksi. Dalam penelitian ini kami mencoba mengeksplorasi teknik pemodelan prediksi kunjungan halaman website yang mampu mengurangi kompleksitas model namun tetap bisa mempertahankan *aplicability* model dan akurasi prediksi. Kami menunjukkan dibandingkan dengan model n-gram, model n-gram+ yang dilengkapi dengan skema support pruning dapat mengurangi ukuran model hingga 75% dan mampu mempertahankan aplicability model dan akurasi prediksinya.

Kata kunci: Web Mining, *n*-gram, Markov Chain, Prediksi

WEB ACCESS PREDICTION USING N-GRAM MODEL

Name	:	Elok Sri Wahyuni
NRP	:	2213206714
Supervisor	:	Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc.

ABSTRACT

Prediction and modeling patterns of user visits a website can be measured its performance with some parameters. Measurement parameters that are often used are the complexity of the model, the ability of the model to make predictions (aplicability) and the prediction accuracy. In this study we tried to explore the predictive modeling techniques visit the website pages that can reduce the complexity of the model, but retaining the aplicability models and prediction accuracy. We show compared with n-gram models, models of n-gram + is equipped with a support scheme pruning can reduce the size of the model up to 75 % and is able to maintain the accuracy aplicability models and predictions.

Key Words: Web Mining, *n*-gram, Markov Chain, Prediksi

KATA PENGANTAR

Dengan mengucap puji syukur kehadirat Allah SWT yang telah melimpahkan segala rahmat dan hidayah-Nya sehingga laporan Tesis yang berjudul Prediksi Kunjungan Halaman Website Dengan N-Gram Model dapat penulis selesaikan.

Laporan ini merupakan salah satu syarat kelulusan pada Program Magister Bidang Keahlian Telematika (CIO) pada Jurusan Teknik Elektro, Fakultas Teknologi Industri pada Institut Teknologi Sepuluh Nopember Surabaya untuk meraih gelar Magister Teknik (MT).

Penulis menyadari bahwa laporan Tesis ini bisa selesai berkat bantuan, dukungan, bimbingan dari berbagai pihak. Pada kesempatan ini, Penulis mengucapkan terima kasih yang sebesar-besarnya kepada :

1. Kementerian Kominfo selaku pemberi beasiswa yang telah memberikan kesempatan pada Penulis untuk melanjutkan pendidikan pada Program Magister
2. Bapak Dr. Ir. Yoyon Kusnendar Suprapto, M.Sc. selaku dosen pembimbing atas segala bentuk bimbingan, arahan, semangat serta kesabaran beliau selama proses penyelesaian laporan Tesis ini.
3. Bapak Bapak Dr. I Ketut Eddy Purnama.,ST.,MT., Bapak Dr. Supeno Mardi S.N.,ST.,MT. dan Bapak Dr. Eko Mulyanto Yuniarno, ST., MT. selaku dosen penguji atas segala saran, kritik dan masukannya demi untuk kesempurnaan laporan Tesis ini.
4. Segenap Dosen dan Civitas akademika Program Studi Magister Jurusan teknik Elektro atas segala bentuk bantuan selama penulis menempuh pendidikan.
5. Pihak Pemerintah Daerah Kab. Jombang yang telah memberikan kesempatan pada Penulis untuk melanjutkan pendidikan pada Program Magister.
6. Semua teman-teman Mahasiswa Program Magister Telematika (CIO) angkatan 2013 dan 2012 atas segala bentuk bantuan, bimbingan dan kebersamaannya dalam menyelesaikan studi ini.
7. Suami tercinta David Hermansyah serta keluarga yang selalu memberikan semangat, doa serta bantuan lain yang tidak bisa penulis tuliskan satu per satu.
8. Semua pihak yang tidak dapat tuliskan satu per satu atas segala bentuk bantuan yang diberikan dengan tulus.

Penulis menyadari bahwa tesis ini masih jauh dari kesempurnaan sehingga diharapkan ada kritikan serta saran yang bersifat membangun demi untuk perbaikan penelitian ini dimasa yang akan datang. Penulis berharap agar tesis ini memberikan

manfaat terutama bagi Pemerintah Kabupaten Bantaeng dalam menerapkan kebijakan pemberian usulan rekomendasi mutasi jabatan struktural kepada pengambil kebijakan lebih lanjut.

Surabaya, Januari 2015

Penulis

Daftar Isi

Daftar Tabel

[*Halaman ini sengaja dikosongkan.*]

Daftar Gambar

[*Halaman ini sengaja dikosongkan.*]

BAB 1

Pendahuluan

1.1 Latar Belakang

Seiring perkembangan teknologi web, semakin banyak pula data yang tersedia bagi pengguna web dan jumlahnya terus bertambah. Data web mencakup spektrum informasi yang sangat luas mulai dari data pemerintahan, data pendidikan, data hiburan dan lain sebagainya. Pada waktu yang bersamaan, jumlah pengguna internet yang memanfaatkan data yang tersedia dalam sistem berbasis web juga meningkat.

Peningkatan jumlah akses pengguna yang mengakibatkan standar kinerja sistem web dalam melayani permintaan pengguna menjadi sangat tinggi. Kinerja sistem web melibatkan baik dari sisi kinerja perangkat lunak maupun perangkat keras (server). Salah satu strategi peningkatan kinerja sistem web adalah dengan menerapkan sistem *web cache*. Teknik *caching* memungkinkan sebuah sistem berbasis web untuk menggunakan ulang informasi yang dibutuhkan dengan cara mengambil informasi yang disimpan dalam sebuah media penyimpanan khusus apabila informasi tersebut pernah digunakan sebelumnya. Dengan demikian secara otomatis waktu yang dibutuhkan oleh sebuah sistem dalam menampilkan informasi akan lebih cepat dibandingkan dengan apabila memproses ulang untuk mendapatkan informasi tersebut.

Beberapa penelitian memanfaatkan model prediksi dengan web log file sebagai data training untuk diterapkan pada algoritma web cache. Dibandingkan dengan algoritma web cache tardisional pemanfaatan model prediksi pada algoritma web cache menghasilkan sistem web cache yang lebih efisien dan adaptif [Chuibi Huang, 2013]. Peneliti lain [Qiang Yang, 2003], menerapkan algoritma prediksi berbasis model n-gram untuk digunakan pada GDSF, sistem pengaturan cache yang sudah umum dikenal.

Dalam beberapa tahun terakhir, masalah pemodelan dan memprediksi perilaku pengguna dalam berselancar pada situs web sendiri telah banyak menarik minat beberapa peneliti karena tidak hanya dapat digunakan untuk meningkatkan hasil sistem web cache, namun juga bisa digunakan untuk merekomendasikan halaman terkait(*recommender system*), meningkatkan kinerja mesin pencari, dan personalisasi website [James Pitkow, 1999].

Model Markov yang digunakan untuk mempelajari dan memahami proses stokastik, terbukti cocok untuk pemodelan dan memprediksi perilaku berselancar pengguna pada situs web. Secara umum, masukan untuk masalah ini adalah urutan halaman web yang diakses oleh pengguna dan tujuannya adalah untuk membangun model Markov yang dapat digunakan untuk memprediksi halaman web yang paling mungkin diakses selanjutnya oleh pengguna website. [Martin Labsky, 2006] membandingkan penggunaan model berbasis aturan dan Markov Model untuk melakukan prediksi pada website eccomerce. Peneliti lain [Ajeetkumar S. Patel, 2014] menggunakan algoritma genetik yang dipadukan dengan model markov dalam membuat sistem prediksi yang didasarkan pada tingkah laku pengguna. Sedangkan [Kurian, 2008] menyajikan model markov dalam bentuk tree dibandingkan model markov dalam bentuk matriks untuk keperluan membuat prediksi kunjungan halaman berikutnya.

Dalam beberapa penelitian disebutkan orde pertama model Markov atau model 1-gram disebutkan sangat tidak akurat dalam memprediksi perilaku berselancar pengguna website, karena model ini hanya mengamati histori terakhir.Jumlah data yang besar sangat memungkinkan besarnya data noise yang bisa menurunkan kinerja model untuk membuat prediksi. Akibatnya, model dengan order tinggi yang sering digunakan. Sayangnya, model order tinggi pun memiliki sejumlah keterbatasan yang terkait dengan sedikitnya data yang tersaji yang mengakibatkan data training tidak mencukupi untuk membuat model yang baik, seperti yang dilakukan [James Pitkow, 1999] yang mengusulkan model *longest repeating subsequence* (LRS). Akibatnya cakupan model dalam membuat prediksi juga menurun dan sehingga kadang-kadang akhirnya menghasilkan akurasi prediksi yang buruk.

Jumlah state yang banyak dapat membatasi *aplicability* model n-gram untuk aplikasi seperti sistem web cache dimana batasan penggunaan memori dan waktu respon sangat penting. Dalam makalah ini kami mengeksplorasi teknik yang mencoba untuk mengurangi besarnya Model namun tetap bisa mempertahankan aplicability dan akurasi model prediksi. Teknik ini menggabungkan pemodelan n-gram+ yang digunakan untuk mengidentifikasi pola berselancar dan memprediksi kunjungan halaman website [Zhong Su, 2000] dengan teknik pruning yang berfungsi mengurangi kompleksitas model [Deshpande and Karypis, 2000].

Adapun Susunan struktur pelaporan pada penelitian ini adalah Bab 2 berisi Tinjauan Pustaka dan Landasan Teori, bab 3 mengenai proses yang dilakukan dalam prediksi dengan n-gram model. Bab 4 membicarakan Analisa hasil dan evaluasi dari pengujian yang dilakukan.dan ditutup dengan bab 5 yang berisi kesimpulan dan pekerjaan selanjutnya.

1.2 Rumusan Permasalahan

Dari uraian tersebut di atas, rumusan masalah dalam penelitian ini adalah peningkatan tingkat *aplicability* dan akurasi model dalam membuat prediksi dibarengi dengan peningkatan jumlah state yang digunakan oleh model. Sementara untuk aplikasi seperti sistem cache, akurasi prediksi menjadi jantung dari pengaturan sistem cache namun membutuhkan model yang tidak besar karena batasan penggunaan memori sangat ketat, sehingga diperlukan model yang tidak terlalu kompleks namun bisa mempertahankan akurasi dan *aplicability* yang tinggi.

1.3 Tujuan Penelitian

Adapun tujuan yang ingin dicapai dalam penelitian ini adalah membangun model untuk prediksi kunjungan halaman website dengan menggunakan pendekatan pemodelan statistik n-gram yang memberikan hasil lebih baik dalam *aplicability*, akurasi dan kompleksitas model dibandingkan dengan model n-gram umumnya.

1.4 Manfaat Penelitian

Model prediksi yang memiliki jumlah state yang tidak banyak namun bisa memberikan akurasi dan *aplicability* model yang baik dapat digunakan dalam algoritma sistem web cache sehingga bisa membantu meningkatkan waktu response sistem berbasis web dalam melayani permintaan pengguna

1.5 Batasan Masalah

1. Data yang di ujicoba diambil dari web server log website pemerintah.
2. Kinerja model diukur dari 3 parameter yaitu aplicability, akurasi dan jumlah state pada model

[*Halaman ini sengaja dikosongkan.*]

BAB 2

Tinjauan Pustaka dan Landasan Teori

2.1 Tinjauan Pustaka

Dalam mendukung penelitian ini, penulis mengacu pada beberapa penelitian sebelumnya yang berkenaan dengan sistem pendukung keputusan.

Penelitian yang dilakukan oleh [Deshpande and Karypis, 2000]. Penelitian ini menerapkan beberapa teknik yang berbeda dan menerapkannya pada model markov yang berbeda-beda untuk menghasilkan model yang memiliki tingkat kompleksitas yang rendah dan meningkatkan akurasi prediksi yang dihasilkan. Model markov dalam penelitian ini digunakan untuk memprediksi teks berikutnya yang diketik oleh pengguna pada aplikasi pengolah kata berdasarkan urutan kata pengguna tersebut pada mas lalu. Model juga diterapkan pada data set yang digunakan untuk memprediksi telepon switch yang dilakukan oleh seorang pengguna.

Penelitian yang dilakukan oleh [James Pitkow, 1999]. Penelitian ini bertujuan untuk mengurangi kompleksitas model prediksi kunjungan web tanpa mengganggu tingkat akurasi model dalam membuat prediksi. Untuk itu peneliti hanya menggunakan urutan kunjungan halaman yang terpanjang dan memiliki frekuensi untuk membangun model. Dengan asumsi urutan kunjungan yang panjang cukup menggambarkan profil pengguna sehingga ketepatan akurasi bisa cukup tinggi. selain itu jumlah urutan panjang yang tidak terlalu banyak bisa menjadikan model yang dihasilkan tidak terlalu besar tapi cukup bisa menggambarkan pola kunjungan. kelemahan dari model ini tidak bisa diterapkan pada website yang pengunjungnya memiliki urutan kunjungan yang pendek-pendek. Karena jumlah data sesuai kriteria yang dimiliki akan sangat sedikit sehingga justru akan menurunkan tingkat akurasi prediksi dari model.

Penelitian yang dilakukan oleh [Ajeetkumar S. Patel, 2014] . Dalam penelitian ini peneliti menggunakan algoritma Genetic dan menggabungkannya dengan Model Markov untuk melakukan prediksi kunjungan halaman web berikutnya. Tujuan yang ingin dicapai oleh pengguna adalah menaikkan tingkat akurasi prediksi dan mendapatkan model yang tidak terlalu kompleks. pendekatan algoritma berbasis genetika juga digunakan untuk meringankan kompleksitas pemodelan sistem yang diusulkan dengan menghasilkan urutan optimal pola kunjungan web dengan mengurangi ukuran ruang pencarian. Sistem yang diusulkan diuji pada standar un-

tuk menganalisis akurasi prediksi.Hasil yang diperoleh dengan mencapai 4% sampai 7% lebih baik dari hasil prediksi menggunakan model markov umumnya.

Penelitian yang dilakukan oleh [Josep Domenech and Pont, 2012]. Peneliti melakukan perbandingan kinerja beberapa algoritma prediksi yang digunakan pada sistem prefetching dan mengelompokkan algoritma-algoritma tersebut berdasarkan cara mendapatkan datanya. Peneliti juga menawarkan algoritma Double Dependency Graph untuk melakukan perbaikan prediksi dan menekankan cara bagaimana memperlakukan obyek terkait dengan halaman website, baik halaman website itu sendiri maupun obyek yang dikandungnya seperti gambar atau file multimedia. Jika pada umumnya algoritma prediksi menghilangkan obyek tambahan tersebut dalam penelitian ini obyek tambahan diperlakukan khusus karena berkaitan dengan karakteristik web. Hasil yang diberikan algoritma mampu mengurangi 15% nilai latency, dengan tidak hanya mengurangi kompleksitas model namun juga waktu yang dibutuhkan prosesor dan memori yang diperlukan.

Penelitian yang dilakukan oleh [Kurian, 2008]. Peneliti menggunakan markov model untuk melakukan prediksi dengan menggunakan bentuk yang berbeda. jika pada umumnya model markov digambarkan dalam bentuk matriks, maka peneliti menggambarkan model markov dalam bentuk pohon state. Dalam penelitian ini, model evolusioner dirancang yang menggunakan fungsi fitness. Fungsi fitness adalah jumlah tertimbang presisi dan tingkat cakupan yang model yang ditawarkan. Hal ini membantu untuk menghasilkan model dengan kompleksitas yang berkurang. Hasil menunjukkan bahwa Model yang ditawarkan mampu melakukan secara konsisten dengan akurasi yang baik pada beberapa log file yang berbeda. Pendekatan evolusioner membantu untuk melatih model untuk membuat prediksi yang sepadan dengan pola web browsing saat ini.

Penelitian yang dilakukan oleh [Ingrid Zukerman, 2001]. Keterbatasan metode representasi pengetahuan tradisional untuk pemodelan perilaku manusia yang kompleks menyebabkan dilakukan penyelidikan model statistik. prediksi statistik model memungkinkan mengantisipasi aspek-aspek tertentu dari perilaku manusia, seperti tujuan, tindakan dan preferensi. Peneliti mengembangkan model dalam konteks enterprise.Kemudian dibuat model dengan dua pendekatan utama dalam pemodlena statistik yaitu pemodelan berbasis konten dan kolaboratif, dan mendiskusikan teknik utama yang digunakan untuk mengembangkan prediksi statistik models. Dalam penelitian ini juga dipertimbangkan persyaratan evaluasi model.

2.2 Landasan Teori

2.2.1 Data Mining

Data Mining adalah kegiatan yang meliputi pengumpulan dan pemakaian data historis untuk menemukan keteraturan, pola atau hubungan dalam set data berukuran besar [Liu, 2006] . Hasil dari proses data mining digunakan untuk memperbaiki pengambilan keputusan di masa depan. Beberapa ahli menganggap *data mining* sebagai istilah lain dari *Knowledge Discovery in Databases (KDD)*. Tujuan dari data mining dapat dibedakan berdasarkan tujuan dari penggunaan sistemnya, yaitu :

1. Deskripsi (*Verification*) Deskripsi terfokus pada menemukan pola-pola yang dimengerti manusia dengan menggunakan data mentah yang ada.
2. Prediksi (*Discovery*) Prediksi melibatkan penggunaan beberapa variabel pada data mentah untuk memprediksi pengetahuan yang tidak diketahui atau belum terjadi dari beberapa variabel yang menarik.

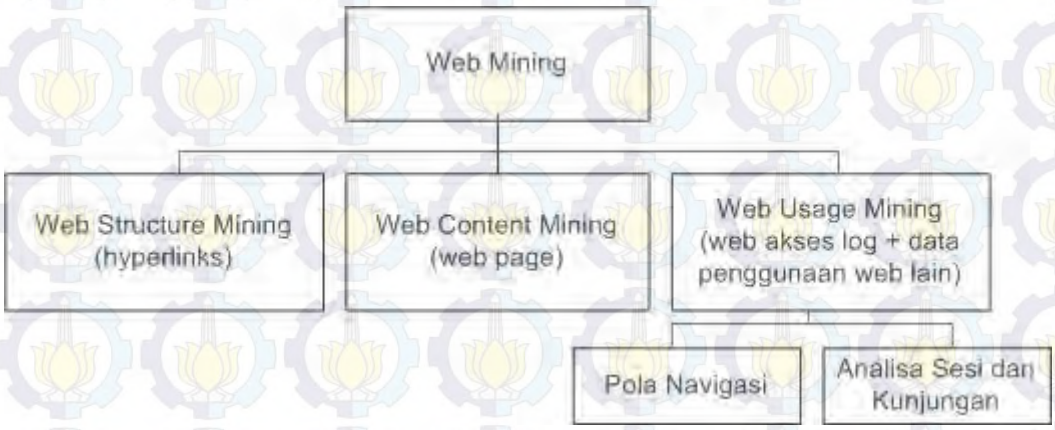
World Wide Web atau lebih dikenal sebagai web adalah sebuah ruang informasi dimana setiap resource didalamnya diidentifikasi dengan suatu pengenal global yang disebut Unified Resource Identifiers (URI). Teknologi web terus berkembang dan kini telah menjadi paradigma yang paling berpengaruh dalam area sistem informasi.

Seiring perkembangan teknologi web, semakin banyak pula data yang tersedia bagi pengguna web dan jumlahnya terus bertambah. data web mencakup spektrum informasi yang sangat luas mulai dari data pemerintahan, data pendidikan, data hiburan dan lain sebagainya.Pada waktu yang bersamaan, semakin banyak pula data yang disimpan dalam penyimpanan tradisional beralih ke web. diprediksikan sebagian besar data yang dibuat manusia disimpan di web.

Namun demikian, ketersediaan data dalam jumlah besar pada web tidak menjamin bahwa seorang pengguna web dapat menemukan data yang dibutuhkannya dengan mudah. Pada kenyataannya, data dalam jumlah besar tersebut, sudah melampaui kemampuan manusia untuk menemukan informasi yang dibutuhkan. namun data web dalam jumlah besar tersebut dengan berbagai macam propertinya menunjukkan bahwa ada pengetahuan tersembunyi dibalik data web yang tidak dapat diinterpretasikan dengan mudah menggunakan intuisi manusia.

Web mining adalah aplikasi dari teknik data mining untuk mendapatkan pola pemakaian pengguna dari data web, sebagai bentuk usaha untuk memahami tingkah laku pengguna web dan memberi servis yang lebih baik pada penggguna

aplikasi berbasis web. Data web mengandung informasi tentang identitas pengguna dengan tingkah laku mereka ketika mengakses sebuah website. Secara umum object data pada web mining adalah resource pada web, dapat berupa dokumen web, log file web, struktur website, atau struktur dari dokumen web itu sendiri. Berdasarkan tipe data yang diolah, web mining dibagi menjadi tiga bidang kajian, yaitu : *Web Content Mining*, *Web Usage Mining*, *Web Structure Mining*. Sehingga taksonomi web mining bisa dilihat seperti pada Gambar 2.1.



Gambar 2.1 Taksonomi Web Mining

Untuk penjelasan masing-masing bidang web mining yang lebih detil akan dijelaskan dalam sub bab-sub bab berikut ini.

2.2.1.1 Web Content Mining

Tujuan dari *web content mining* adalah menemukan pengetahuan yang berguna dengan menggunakan data berupa konten/isi dari dokumen web.

Sebagai contoh, kita bisa secara otomatis mengkalsifikasikan atau mengklaster halaman-halaman web berdasarkan topiknya. Sekilas teknik ini hampir sama seperti pada penambangan data tradisional. Namun kita juga bisa menemukan pola pada halaman-halaman web untuk mendapatkan data yang berguna seperti deskripsi produk, posting yang masuk disebuah forum, dan lain-lain untuk berbagai tujuan. selanjutnya kita bisa menambang review pembeli dan posting forum untuk mendapatkan sentimen pembeli. Proses seperti ini tidak bisa dilakukan pada penambangan data tradisional.

Salah satu aplikasi dari web content mining yang banyak digunakan adalah search engine.

2.2.1.2 Web Structure Mining

Web Structure Mining adalah proses mendapatkan pengetahuan dari topologi sebuah web yang mencakup link-link diantara halaman-halaman web yang merepresentasikan struktur sebuah website.

Pada web structure mining ini dapat dianalisa halaman mana saja yang menjadi target link dari halaman web lainnya, halaman web mana yang mengacu ke banyak halaman web lainnya, atau koleksi halaman mana yang membentuk sebuah kumpulan.

Dari hyperlink (atau link) kita bisa menemukan halaman-halaman mana yang paling penting, dimana hal ini merupakan sebuah kunci dari tekhnologi mesin pencari. kita juga bisa menemukan komunitas dari pengguna yang membagikan hal-hal yang menjadi ketertarikan mereka. Penambangan data tradisional tidak bisa melakukan hal seperti ini karena tidak ada struktur link pada database relational

2.2.1.3 Web Usage Mining

Tujuan dari web usage mining adalah untuk mengambil, memodelkan dan menganalisa pola tingkah laku dan profil pengguna yang berinteraksi dengan sebuah website. Pola yang dihasilkan biasanya berbentuk kumpulan halaman, objek atau sumber daya lain yang sering diakses oleh grup penggguna yang sama kebutuhan atau ketertarikannya.

Web usage mining mempelajari perilaku, pola kunjungan dan data yang berhubungan yang dihasilkan dari interaksi pengguna dengan satu atau lebih website. Analisa tingkah laku tersebut bisa dalam jangka waktu sebuah sesi atau juga bisa mencapai beberapa tahun.

Secara garis besar proses web usage mining terdiri dari tiga fase yaitu :

1. pengumpulan data dan praproses
2. pengenalan pola
3. menganalisa pola

Pengumpulan Data dan Praproses

Pada tahap praproses, data /textitclick-stream dibersihkan dan dibagi menjadi sejumlah transaksi pengguna yang menggambarkan aktifitas pengguna selama berkunjung ke sebuah website. Sumber lain seperti struktur website atau isi web juga bisa digunakan pada tahap pra proses ini. Pra proses memegang peranan penting dalam web usage mining, karena hasil representasi data dari pra proses bisa

mempengaruhi hasil penambangan data secara keseluruhan. oleh karena itu banyak penelitian yang dikembangkan untuk mendapatkan teknik pra proses yang efektif dan benar - benar bisa menggambarkan data yang dibutuhkan dalam proses selanjutnya.

Sebuah langkah penting pada setiap aplikasi data mining adalah pembentukan data set yang sesuai dengan target dimana data mining dan algoritma statistik akan diaplikasikan. Hal ini juga berlaku pada teknik web usage mining, terkait dengan karakteristik data urutan aksi yang dipakai dan hubungannya dengan data lain yang dikumpulkan dari beberapa sumber yang berbeda. Sehingga untuk memproses data tersebut dibutuhkan algoritma khusus yang tidak bisa dipakai pada bidang lain. Proses tersebut bisa meliputi praproses data asli, menggabungkan data dari berbagai sumber, mengubah gabungan data tersebut menjadi bentuk yang sesuai untuk masukan bagi operasi data mnining yang spesifik. Secara umum proses - proses ini disebut sebagai proses persiapan data/praproses.

Elemen penting pada praproses meliputi :

1. Data Fusion dan Pembersihan data Pada website yang berskala besar, konten yang disajikan pada pengguna berasal dari beberapa website atau server aplikasi. pada beberapa kasus, beberapa server dengan data yang sama digunakan untuk mengurangi beban kerja server tertentu. Data fusion digunakan untuk menggabungkan log file darai beberapa website atau server aplikasi.
   Pembersihan data biasanya bersifat khusus bergantung pada website dimana data diambil, karena melibatkan proses seperti penghapusan informasi tambahan yang mungkin tidak diperlukan untuk analisa yang dilakukan seperti file gambar, atau multimedia lain.
2. Identifikasi Pageview sebuah pageview secara konsep setiap pageview adalah sebuah kumpulan obyek website yang merepresentasikan sebuah ”user event” yang spesifik seperti membaca sebuah artikel, melihat halaman atau memasukkan produk pada shopping chart. Identifikasi pageview tergantung pada struktur hubungan antar halaman.
3. Identifikasi Pengguna Analisa web usage tidak membutuhkan pengetahuan tentang identitas pengguna. Namun sangat penting untuk membedakan pengguna. Karena seorang pengguna mungkin mengunjungi sebuah website lebih dari satu kali, log server mencatat beberapa sesi dari setiap pengguna.
4. Sessionization Sessionization adalah proses segmentasi record aktifitas pengguna dari setiap pengguna menjadi beberapa sesi, dimana setiap sesi merepresentasikan satu kali kunjungan ke website.
5. Path Completion sistem cache proxy seringkali mengakibatkan hilangnya re-

ferensi ke sebuah halaman yang datanya tersimpan di cache. Referensi yang hilang karena proses cache secara heuristic bisa dilengkapi dengan menggunakan proses path completion.

6. Integrasi Data hasil dari praproses diatas adalah sekumpulan sesi pengguna dimana setiap sesi pengguna adalah urutan terbatas halaman yang dilihat/diakses pengguna. untuk lebih meningkatkan efektifitas proses penemuan pola, data dari berbagai sumber harus disatukan sebelum dilakukan preproses.

Sumber data utama untuk web usage mining adalah file server log yang terdiri dari web server access log dan server log aplikasi. Sumber data tambahan yang juga penting baik untuk pra proses maupun fase pengenalan pola adalah file situs dan meta data, database operasional, template aplikasi dan pengetahuan utama. Untuk beberapa kasus dan beberapa user, data tambahan mungkin tersedia dari sisi klien maupun dari proxy server.

Data yang digunakan pada web usage mining berdasarkan sumber data tersebut diambil, dibagi menjadi 4 kategori :

1. *Usage Data*. Usage data yang paling mudah diperoleh untuk keperluan web usage mining adalah web log. web access log adalah sebuah log yang mencatat semua tingkah laku pengunjung sebuah website, misalnya IP Address pengunjung, waktu kunjungan, atau halaman mana saja yang dikunjungi. Setiap hit yang diterima server, berhubungan dengan permintaan HTTP, membangkitkan sebuah entri tunggal di akses log server. Setiap log entry bisa mengandung beberapa field yang mengidentifikasikan waktu dan jam request, alamat IP klien, status request, metode HTTP yang digunakan, user agen (seperti browser, jenis dan tipe OS), web yang mereferensi bahkan bisa juga cokies dari klien.
2. *Content Data* Data konten pada sebuah website adalah kumpulan obyek dan hubungan yang disajikan kepada pengguna. secara umum data jenis ini adalah kombinasi dari materi bebrbentuk teks dan gambar. data sumber yang digunakan untuk mengirim atau membangun data ini adalah halaman HTML/XML , file multimedia , bagian halaman yang dibangun dari script dan kumpulan record dari operasi database. selain itu yang termasuk data konten website adalah semantic or structural meta-data yang menempel pada pada website atau setiap halaman, seperti descriptive keywords, document attributes, semantic tags, or HTTP variables.
3. *Structure Data* Data struktur merepresentasikan pandangan designer web dalam mengatur isi dari sebuah website. Pengaturan ini bisa dilihat dari hubungan antar halaman yang direfleksikan sebagai *hyperlink* pada website tersebut.

4. *User Data* Operasi database pada sebuah website mengandung informasi tambahan dari profil pengguna website. Data tersebut bisa mengandung informasi demografis tentang pengguna yang mendaftar, rating pengguna terhadap bebagai obyek seperti produk atau film, catatan pembelian atau sejarah kunjungan pengguna yang semuanya bisa menggambarkan bentuk ketertarikan pengguna.

Pemodelan Data untuk Web Usage Mining

Hasil dari praproses adalah sekumpulan n kunjungan halaman, P=p1, p2, , pn, dan sekumpulan m transaksi, T= t1,t2,,tm, dimana setiap ti pada T adalah sebuah subset dari P. Pageview adalah urutan entitas bermakna dimana fungsi penambangan dilakukan (seperti halaman web atau produk). Secara konsep setiap transaksi t dilihat sebagai urutan dengan panjang l dari pasangan berurut :

$$t = (p1^t, w(p1^t)), (p2^t, w(p2^t))...(pl^t, w(pl^t)) \quad (2.1)$$

dimana setiap $p_i^t = p_j$ dimana nilai j adalah 1,2...,n dan $w(p1^t)$ adalah bobot dari pageview p_i^t pada transaksi t.

Bobot bisa ditentukan dalam beberapa cara, bergantung pada tipe analisa atau personalisasi framework. Sebagai contoh, pada aplikasi collaborative filtering yang sangat bergantung pada profil pengguna yang sama untk membuat rekomendasi kepada pengguna sekarang, bobot didasarkan pada rating pengguna dari item. Pada sebagian besar web usage mining, bobot bernilai biner, merepresentasikan pageview yang muncul maupun yang tidak muncul pada transaksi atau bobot bisa juga adalah sebuah fungsi dari durasi pageview pada sesi pengguna.

Pengenalan dan Analisa Pola

Pada tahap pengenalan pola statistik, database atau operasi *machine learning* digunakan untuk mendapatkan pola tertentu dari tingkah laku pengguna website selama berinteraksi dengan website. Kemudian pola dan statistik yang didapatkan selanjutnya diproses atau difilter untuk menghasilkan model pengguna akhir sebagai input untuk aplikasi lain seperti tool visualisasi, sistem pendukung keputusan dan lain-lain.

Tipe dan level analisa yang diterapkan pada usage data bergantung tujuan khusus dari analisa tersebut dilakukan dan keluaran apa yang diharapkan. Berikut ini adalah beberapa tipe pengenalan pola dan teknik analisa yang biasa digunakan

pada web usage mining :

1. Session and Visitor Analysis

   Analisa statistik dari praproses data sesi nengandung bentuk analisis yang umum. dalam kasus ini data adalah didefinisikan dengan satuan unit seperti hari, sesi, jumlah pengunjung atau domain. teknik statistik standar bisa digunakan pada data ini untuk mendapatkan pengetahuan tentang tingkah laku berkunjung. pendekatan ini banyak diaplikasikan oleh sebagian besar tool komersil untuk analisa web log. bentuk lain untuk analisa dan integrasi data usage adalha Online Analytical Processing (OLAP). OLAP menawarkan framework yang terintegrasi untuk analisa dengan derajat keluwesan yang lebih tinggi/ lebih fleksibel. Another form of analysis on integrated usage data is Online Dta masukan untuk analisa OLAP adalah data multidimensi yang menggabungkan penggunaan, konten, data e-commerce dalam berbagai level dari integrasi untuk setiap dimensi.

2. Cluster Analysis and Visitor Segmentation

   Pengklasteran adalah teknik penambangan data yang bekerja dengan mengelompokkan sebuah set item yang memiliki kesamaan karakteristik. Pada bidang web usage mining ada dua klaster yang bisa ditemukan yaitu klaster pengguna dan klaster halaman. Klaster record pengguna (sesi dan transaksi) adalah yang paling banyak digunakan untuk analisa pada web usage mining dan analisa web.

   Pengelompokan pengguna cenderung membentuk kelompok-kelompok pengguna yang menunjukkan pola berselancar yang sama. Pengetahuan semacam ini sangat berguna untuk menyimpulkan demografi pengguna dalam rangka untuk melakukan segmentasi pasar apliaksi e-commerce atau untuk personalisasi web yang memberikan konten kepada pengguna dengan ketertarikan yang sama.

   Pengklasteran halaman (atau item) dapat dilakukan berdasarkan data penggunaan (yaitu, mulai dari pengguna sesi atau data transaksi), atau berdasarkan fitur konten yang berhubungan dengan halaman atau item (kata kunci atau atribut produk). Dalam kasus pengklasteran berbasis konten, hasilnya mungkin berupa koleksi halaman atau produk yang berhubungan dengan topik atau kategori yang sama. Dalam pengelompokan berbasis penggunaan berbasis, item yang sering diakses atau dibeli bersama-sama dapat secara otomatis diatur ke dalam kelompok-kelompok. Hal ini juga dapat digunakan untuk menyediakan halaman HTML permanen atau dinamis yang menyarankan hyperlink terkait dengan pengguna sesuai dengan sejarah masa lalu mereka dalam

melakukan pembelian.

3. Association and Correlation Analysis
   Penemuan dengan Aturan asosiasi dan analisis korelasi statistik dapat menemukan kelompok item atau halaman yang sering diakses atau dibeli bersama-sama. Hal ini, pada gilirannya, memungkinkan situs web untuk mengatur isi situs lebih efisien, atau memberikan rekomendasi produk cross-sale yang efektif . Pendekatan yang paling umum untuk penemuan asosiasi didasarkan pada algoritma Ap-riori. Algoritma ini menemukan kelompok barang (page-view yang muncul di log hasil praproses) yang sering terjadi muncul bersama-sama dalam banyak transaksi.
4. Analysis of Sequential and Navigational Patterns
   Teknik sequential pattern mining mencoba untuk menemukan pola antar-sesi seperti kehadiran satu set item diikuti oleh item lain dalam waktu/sesi/episode yang ditetapkan. Dengan menggunakan pendekatan ini, Web pemasar dapat memprediksi pola kunjungan masa depan yang akan membantu dalam menempatkan iklan yang ditujukan untuk kelompok pengguna tertentu.
5. Classification and Prediction based on Web User Transactions
   Klasifikasi adalah tugas pemetaan item data ke dalam salah satu dari beberapa kelas yang ditentukan. Dalam domain Web, salah satu yang tertarik untuk mengembangkan profil pengguna milik kelas tertentu atau kategori. Hal ini memerlukan ekstraksi dan pemilihan fitur yang paling menggambarkan sifat-sifat yang diberikan kelas atau kategori. Klasifikasi dapat dilakukan dengan menggunakan supervised learning algoritma seperti pohon keputusan, pengklasifikasi Naif Bayesian , k-nearest neighbor dan Support Vector Machines

2.2.2 Pemodelan Statistik

Sebuah model statistik adalah seperangkat asumsi tentang pembentukan data yang diamati, dan data yang mirip dari populasi yang lebih besar. Sebuah model merupakan bentuk ideal yang diharapkan dihasilkan dari proses pembentukan data. Sebuah model statistik digunakan untuk menggambarkan distribusi probabilitas, dimana beberapa diantaranya diasumsikan cukup mendekati dsitribusi sebenarnya dari data sampel. Model ditetapkan oleh hubungan dari sebuah atau beberapa variabel acak dengan variabel-variabel tidak acak. Hubungan ini biasanya diberikan dalam persamaam matematika. Mengutip perkataan Kenneth Bollen, Sebuah model adalah representasi formal dari teori. Semua uji statistik dapat digambarkan dalam model statistik. sebagai contoh, misalkan tes untuk membandingkan rata-rata dari

dua set data dapat dirumuskan dengan menentukan jika sebuah paramater yang diestimasikan pada model bernilai selain dari 0. Semua perkiraan statistik diturunkan dari model statistik.

Dalam kalimat matematika, model statistik biasanya dianggap sebagai sepasang (S, P), dimana S adalah himpunan kemungkinan pengamatan, yaitu ruang sampel, dan P adalah satu set distribusi probabilitas pada S.

Contoh model statistik diilustrasikan dalam contoh berikut. umur dan tinggi badan manusia masing - masing dapat didistribusikan secara probabilistik pada manusia. Umur dan tinggi secara stokastik berkaitan. jika seseorang berumur 10, maka akan mempengaruhi kemungkinan orang tersebut memiliki tinggi badan 100 m. maka hubungan ini bisa diformulasikan dalam model regeresi linier dalam rumus berikut :

$$height_i = b_0 + b_i age_i + _i \tag{2.2}$$

,dimana b_0 adalah intercept, dan b_i adalah sebuah parameter yang dikalikan age untuk mendapatkan prediksi nilai height, adalah variabel error dan i adalah identitas orang. artinya height (tinggi) diprediksikan oleh age (usia), dengan beberapa nilai error.

Sebuah Model harus cocok untuk semua titik data. Dengan demikian, garis lurus

$$height_i = b_0 + b_i age_i \tag{2.3}$$

bukanlah model data. Garis tersebut tidak bisa menjadi model, kecuali persis cocok semua poin yaitu data di semua titik yangterletak sempurna pada garis lurus. Istilah kesalahan, i, harus dimasukkan dalam model, sehingga model dapat menjelaskan semua titik data.

Ada tiga tujuan untuk model statistik.[Konishi S., 2008]

1. Prediksi
2. Ekstraksi informasi
3. Deskripsi struktur stokastik

Pemodelan statistik Untuk Akses Web dilakukan dengan melakukan memodelkan tingkah laku pengguna yang nantinya digunakan untuk menyimpulkan informasi yang tidak teramati dengan mengamati informasi yang bisa diamati dari pengguna, misalnya tindakan atau ucapan pengguna web.

Pada masa awalnya, sistem pemodelan pengguna dibangun berbasis pengetahuan yang dibentuk untuk membentuk kesimpulan dari pengamatan tentang pengguna. Basis pengetahuan ini biasanya dibangun dengan menganalisis beberapa contoh masalah yang dihadapi yang dianggap bisa mewakili masalah ini. Na-

mun, basis pengetahuan ini mempunyai kelemahan yaitu konstruksinya sarat dengan proses yang intensif dan biasanya tidak mudah beradaptasi dan sulit dikembangkan.[Ingrid Zukerman, 2001]

Model statistik dibangun sebagai alternatif dari model tradisional dalam pemodelan pengguna sebagai jawaban untuk mengatasi masalah pada pemodelan tradisional. Kecerdasan Buatan sebagai area dari mesin pembelajaran dan penalaran dari ketidakpastian telah menghasilkan berbagai teknik yang didasarkan dari model prediksi statistik seperti decision tree, jaringan saraf dan bayesian.

Prediksi dengan menggunakan teknik ini telah digunakan untuk mengadaptasi perilaku sistem. Dua pendekatan utama telah diadopsi untuk melakukan tugas prediksi yaitu content base / berbasisi konten dan kolaboratif. Pembentukannya didasarkan pada prinsip bahwa setiap penggguna menunjukkan tingkah laku tertentu dibawah sejumlah keadaan, dan bahwa perilaku ini diulang dibawah kondisi serupa. Selanjutnya didasarkan prinsip bahwa orang-orang dalam kelompok tertentu cenderung berperilaku sama dibawah sejumlah kondisi tertentu. Dengan demikian, dalam pendekatan berbasis konten, perilaku pengguna diperkirakan dari perilakunya di masa lalunya, sedangkan pada pendekatan kolaboratif, perilaku pengguna diperkirakan dari perilaku orang lain yang berpikiran sama.

Secara umum,Isu yang penting pada sistem pemodelan pengguna dan model prediksi dengan statistik khususnya adalah berkaitan dengan evaluasi dari sistem ini. Sistem prediksi statistik untuk pemodelan pengguna mewarisi aturan persyaratan dan evaluasi dari dua disiplin ilmu yaitu mesin pembelajaran dan pemodelan sistem. Umumnya evaluasi mesin pembelajaran terdiri dari membagi data set menjadi training set dan set tes, menggunakan algoritma untuk mengajari model dan selanjutnya mengevaluasi performa dari model. Metodologi ini sudah diaplikasikan pada pemodelan pengguna dengan statistik yang dikembangkan sampai saat ini dengan menggunakan beberapa alat ukur : recall dan precision, probabilitas prediksi dan akurasi dan utility. Berbeda dengan evaluasi mesin pembelajarann saat ini tidak ada metodologi yang berlaku secara umum untuk mengevaluasi sistem yang menggunakan pemodelan pengguna.

Secara umum ada dua pendekatan utama diadopsi untuk membangun Model statistik untuk prediksi, yaitu :

1. Model statistik berbasis konten.
2. Model statistik kolaboratif

Untuk Pemodelan Berbasis konten, pembelajaran yang digunakan saat pengguna perilaku masa lalu merupakan indikator yang dapat diandalkan / perilaku masa depan nya. Dalam pendekatan ini, model prediksi yang dibangun untuk pengguna

yang menggunakan Data dari / perilaku masa lalu nya. Model berbasis konten sangat cocok untuk situasi di mana pengguna cenderung menunjukkan perilaku aneh. Namun, pendekatan ini memerlukan sistem untuk mengumpulkan data dalam jumlah yang relatif besar dari setiap pengguna untuk memungkinkan dilakukan model statistik.

Pembelajaran kolaboratif ini digunakan setiap kali seseorang dapat berasumsi bahwa pengguna berperilaku dalam cara yang mirip dengan pengguna lain. Dalam pendekatan ini, model dibangun dengan menggunakan data dari kelompok pengguna, dan kemudian digunakan untuk membuat prediksi tentang pengguna individu.

Beberapa model statistik baik yang berbasis konten maupun dengan pendekatan kolaboratif telah banyak digunakan. Model model yang umum antara lain : linear models, TFIDF-based models, Markov models, neural networks, klasifikasi and rule-induction methods dan Bayesian networks.

Model linier memiliki struktur yang sederhana, yang membuat model linear mudah dipelajari, dan juga memungkinkan untuk diperluas dengan mudah. Model linier menggunakan sejumlah dari nilai yang dikenal untuk menghasilkan nilai untuk kuantitas yang tidak diketahui. Sebagai contoh, dengan menggunakan pendekatan kolaboratif untuk membangun sebuah model linier yang memprediksi Peringkat pengguna untuk artikel berita. Dalam model ini, untuk setiap artikel kandidat, untuk nilai yang diketahui misalkan rating artikel, diberikan oleh user lain ke artikel tersebut, dan bobot diukur dari kemiripan antara pertanyaan yang dinberikan oleh satu user dengan user ynag lain. Hasil dari model linear adalah bobot dari total rating. Model linear juga digunakan pada pndekatan berbasis konten, misalnya prediksi waktu antara seorang berhasil login atau memprediksi rating film oleh pengguna.

Metode TFIDF (Term Frequency Inverse Document Frequency) adalah sebuah skema pembobotan yang biasa digunakan dalam bidang mendapatkan kembali informasi dan dokumen yang sesuai dengan pencarian pengguna (Salton dan McGill, 1983). Metode ini merepresentasikan dokumen sebagai vektor bobot, di mana masing-masing bobot sesuai dengan istilah dalam dokumen.Beberapa penelitian menerapkan model TFIDF pada sistem berbasis konten yang merekomendasikan dokumen ke pengguna berdasarkan dokumen lain yang mirip yang menarik ke pengguna ini. Pengembangan pendekatan ini dengan menggunakan genetik algoritma untuk secara otomatis menyesuaikan sistem rekomendasi untuk pengguna

Seperti model linear, model Markov memiliki struktur sederhana. Hal ini disebabkan mereka ketergantungan pada Markov asumsi yang mewakili urutan per-

istiwa (menurut asumsi ini, terjadinya acara berikutnya hanya bergantung pada sejumlah yang tetap dari peristiwa sebelumnya). Dari sejumlah kejadian yang diamati, kejadian berikutnya diperkirakan dari distribusi probabilitas dari peristiwa yang mengikuti kejadian yang diamati dimasa lalu tersebut. Misalnya, untuk masalah memprediksi Halaman web yang akan dikunjungi oleh pengguna, kejadian terakhir yang diamati bisa jadi adalah halaman terakhir yang dikunjungi oleh pengguna atau yang mengandung informasi tambahan seperti ukuran dokumen.Bestavros (1996), Horvitz (1998) dan Zukerman et al. (1999) menggunakan model Markov dengan pendekatan kolaboratif untuk memprediksi permintaan pengguna di web. Model Bestavros menghitung probabilitas bahwa pengguna akan meminta dokumen tertentu di masa depan. Model Horvitz menghitung probabilitas bahwa pengguna akan meminta dokumen tertentu dalam permintaan berikut, dan Zukerman et al. (1999) membandingkan kinerja prediksi model Markov yang berbeda untuk memprediksi kunjungan berikutnya pada web. Prediksi yang dihasilkan oleh model ini kemudian digunakan oleh sistem presending dokumen yang mirip dengan yang diminta oleh pengguna atau bagian dari dokumen-dokumen tersebut.

Jaringan saraf mampu mengungkapkan beragam keputusan non-linear. Hal ini dilakukan melalui struktur jaringan, batas non-linear dan bobot dari tepi antara node. Jennings dan Higuchi (1993) menggunakan jaringan saraf dengan pendekatan berbasis konten untuk mewakili preferensi pengguna untuk artikel berita. Untuk setiap pengguna, mereka belajar jaringan saraf di mana node merepresentasikan kata-kata yang muncul di beberapa artikel yang disukai oleh pengguna dan ujung-ujungnya merepresentasikan kekuatan hubungan antara kata-kata yang muncul dalam artikel yang sama.

Metode klasifikasi mempartisi satu set objek ke dalam kelas sesuai dengan nilai atribut dari objek objek tersebut. Diberikan ruang n-dimensi yang sesuai dengan atribut yang dipertimbangkan, cluster atau kelas yang dihasilkan berisi item yang berdekatan satu sama lain dalam ruang tersebut dan memiliki jarak yang jauh dari klaster lainnya. Metode klasifikasi adalah metode unsupervised. Melalui pendekatan kolaboratif, Perkowitz dan Etzioni (2000) menggunakan variasi pengklasteran tradisional untuk secara otomatis membuat halaman indeks yang berisi link ke halaman web yang terkait satu sama lain (ini adalah halaman yang pengguna cenderung untuk mengunjungi selama sesi yang sama). Teknik klasifikasi, yang mereka sebut clustermining, adalah sejumlah kecil klaster berkualitas tinggi (bukan partisi seluruh ruang dokumen) dimana beberapa dokumen bisa berada pada beberapa klaster yang berbeda.

Rule induction terdiri dari pembelajaran seperangkat peraturan yang memp-

rediksi kelas dari observasi terhadap aribut-atribut. Teknik yang digunakan untuk aturan induksi berbeda dari yang digunakan untuk klasifikasikasi dimana bahwa selama pelatihan, teknik rule induction membutuhkan kelas masing-masing pengamatan beserta atributnya. Model yang diperoleh dari teknik ini dapat mewakili aturan secara langsung, atau mewakili aturan sebagai pohon keputusan atau pada bidang probabilitas bersyarat. Billsus dan Pazzani (1999) menerapkan campuran metode rule induction dan model TFIDF berbasis dan linear untuk merekomendasikan artikel berita kepada pengguna. Sistem mereka menggunakan dua model untuk mengantisipasi apakah pengguna akan tertarik dalam sebuah artikel kandidat. teknik ini sangat berguna ketika membangun sebuah model awal berdasarkan data yang terbatas, karena hanya beberapa artikel berita yang diperlukan untuk mengidentifikasi kemungkinan topik yang menarik. Model lain menerapkan klasifikasi Naiv Bayes (Duda dan Hart, 1973) untuk merepresentasikan fitur vektor Boolean dari artikel kandidat, di mana masing-masing fitur menunjukkan ada atau tidak adanya kata dalam artikel. Melalui pendekatan kolaboratif, Basu et al. (1998) menggunakan Ripper untuk proses pembelajaran satu set aturan yang memprediksi apakah pengguna akan suka atau tidak suka suatu lm , dan Litman dan Pan (2000) digunakan Ripper untuk mempelajari seperangkat aturan yang mengadaptasi strategi dialog yang digunakan dengan sistem dialog yang diucapkan. Gervasio et al. (1998) menggunakan ID3 (Quinlan, 1986) untuk belajar pohon keputusan yang memprediksi mana tindakan yang akan dilakukan selanjutnya oleh pengguna bekerja pada masalah penjadwalan.

Jaringan Bayesian (Bayesian networks) (Pearl, 1988) dan berbagai pengembangan dari BNs sangat populer di komunitas kecerdasan buatan dan telah digunakan untuk berbagai pemodelan pengguna (Jameson, 1995). Jaringan Bayes digambarkan dengan acyclic graph, di mana node adalah variabel acak. Node dihubungkan dengan garis yang menggambarkan sebagai hubungan sebab akibat dari node induk untuk anak-anak mereka. Setiap node dikaitkan dengan distribusi probabilitas bersyarat yang memberikan probabilitas untuk setiap nilai kemungkinan node tersebut untuk setiap kombinasi dari nilai-nilai node induknya. BNS lebih fleksibel daripada model yang lainnya, dalam arti bahwa bayesian memberikan representasi kompak dari setiap distribusi probabilitas, dan secara eksplisit merupakan hubungan sebab akibat, dan bayesian memungkinkan prediksi yang akan dibuat menggunakan sejumlah variabel (bukan variabel tunggal, yang biasa digunakan pada model model yang lain).

Properti penting dari bayesian adalah bayesian memungkinkan kombinasi dari pendekatan kolaboratif dan pendekatan berbasis konten. Pendekatan Kolabo-

ratif digunakan untuk mendapatkan tabel probabilitas bersyarat dan keyakinan awal dari Bayesian. keyakinan ini kemudian dapat diperbarui pada aturan dari basis konten ketika jaringan diakses oleh pengguna. Mode operasinya memungkinkan model prediktif untuk mengatasi masalah pengoleksian data pada pendekatan berbasis konten (yang membutuhkan data dalam jumlah besar yang akan dikumpulkan dari satu pengguna), sementara pada saat yang sama memungkinkan merangkai dari aspek pembelajaran model kolaboratif untuk pengguna tunggal. BNS dan ekstensi mereka telah digunakan untuk melakukan berbagai tugas prediktif. Horvitz et al. (1998) menggunakan BN untuk memprediksi jenis bantuan yang dibutuhkan oleh pengguna melakukan tugas-tugas spreadsheet; Lau dan Horvitz (1999) membangun sebuah model bayesian untuk query pencarian di web dan memprediksi jenis tindakan dan permintaan berikutnya terkait dengan pengguna tersebut. Albrecht et al. (1998) membandingkan kinerja beberapa jaringan Bayesian dinamis yang memprediksi tindakan berikutnya pengguna.

Horvitz et al. (1999) menggunakan jaringan Bayesian dinamis untuk memprediksi perhatian pengguna dan jarak waktu antara pengecekan email yang berturut-turut. Prediksi ini digunakan pada sistem yang memutuskan pemberian peringatan jika ada email yang masuk dan bagaimana melakukannya.

2.2.3 Markov Model

Teori probabilitas modern mempelajari proses kebetulan dimana pengetahuan tentang hasil sebelumnya mempengaruhi prediksi untuk percobaan berikutnya. Pada prinsipnya, ketika kita mengamati urutan eksperimen kesempatan, semua hasil masa lalu bisa mempengaruhi prediksi untuk percobaan berikutnya. Sebagai contoh, untuk kasus memprediksi nilai siswa pada urutan ujian dalam suatu kursus. Pada tahun 1907, A.A. Markov mulai mempelajari proses kesempatan jenis baru yang penting. Dalam proses ini, hasil percobaan yang diberikan dapat mempengaruhi hasil percobaan berikutnya. Jenis proses ini disebut rantai Markov.

Kami menjelaskan rantai Markov sebagai berikut: Kami memiliki satu set states, S = (s1, s2,...sr).Proses dimulai di salah satu state ini dan bergerak berturut-turut dari satu state ke state yang lain. Setiap langkah disebut step. Jika rantai saat ini dalam state si, maka langkah menuju state sj, dengan probabilitas yang dilambangkan pij, dan probabilitas ini tidak tergantung pada yang menyatakan rantai semula sebelum state saat ini. Probabilitas pij disebut probabilitas transisi. The probabilitiespij are calledtransition probabilities. proses bisa berputar pada state yang sama dan langkahnya memiliki nilai probabilitas pii. Probabilitas awal distribusi,

didefinisikan pada S, yang menentukan keadaan awal. Biasanya ini dilakukan oleh state yang disebut sebagai state awal.

Rantai markov digunakan untuk menggambarkan urutan halaman-halamn yang diakses oleh pengguna website, dan digunakan untuk memprediksi halaman berikutnya yang akan diakses oleh pengguna. Model ini dikenal dengan sebutan Model Markov atau Model n-gram Markov.

Misalkan P = p1, p2, . . . , pm adalah kumpulan halaman pada sebuah website. Dan W adalah sebuah sesi pengguna yang mengandung urutan halaman yang diakses pengguna dalam sebuah kunjungan. Diasumsikan bahwa pengguna mengunjungi l halaman, maka prob(pi—W) adlah probabilitas pengguna akan mengakses halaman pi. Halaman pl+1 dimana pengguna akan mengunjungi diestimasikan oleh :

$$Pl+1 = argmaxpIPP(Pl+1 = p|W) = argmaxpIPP(Pl+1 = p|pl, pl1, ..., p1) \tag{2.4}$$

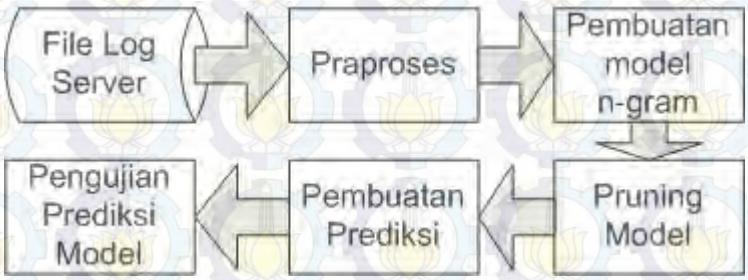
probabilitas prob(pi—W) diestimasi dengan menggunakan semua urutan W oleh semua pengguna pada data training, dinotasikan dengan W. Umumnya semakin panjang l dan semakin besar W, maka semakin akurat prob(pi—W). Namun sangat sulit mendapatkan l yang sangat panjang dan W yang sanagt besar, dan biasanya justru mengarah pada kompleksitas yang tidak perlu. Maka digunakan berbagai model Markov yang dimodifikasi.

[*Halaman ini sengaja dikosongkan.*]

BAB 3

Metodologi Penelitian

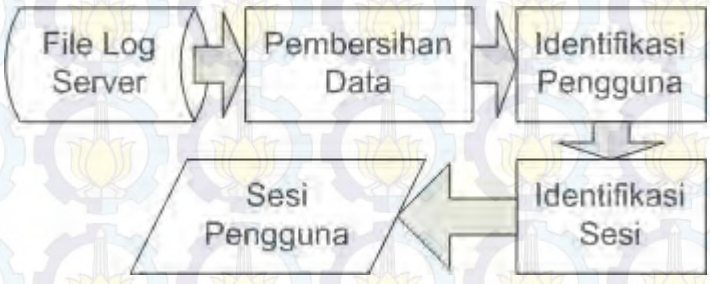
Prediksi kunjungan halaman website dengan model n-gram ini dibangun dari log file web server. Log File jenis ini mencatat semua permintaan halaman web yang dilakukan oleh pengunjung sebuah website. Adapun tahapan proses yang dilakukan untuk membuat model n-gram dan prediksi kunjungan halaman berikutnya ini ditunjukkan dalam blok diagram pada Gambar 3.1



Gambar 3.1 Blok Diagram Penelitian

Blok diagram tersebut dibuat untuk memudahkan dalam memetakan langkah-langkah yang akan dilakukan dalam penelitian untuk pembuatan model dan melakukan evaluasi prediksi yang dihasilkan. Langkah langkah tersebut seperti pada Gambar 3.1 di atas akan dijabarkan lebih lanjut sebagai berikut :

3.1 Praproses



Gambar 3.2 Tahapan Praproses

Tujuan dari praproses adalah mengolah data mentah berupa file log web server menjadi data yang siap digunakan dalam pemodelan dalam bentuk kumpulan

sesi pengguna. Masukan/keluaran dan tahapan pada praproses seperti diilustrasikan pada Gambar 3.2. Masing masing proses akan dijelaskan sebagai berikut :

3.1.1 Pembersihan Data

Tabel 3.1 Detil Apache Combined Log Format

Nama Field	Deskripsi	Contoh Nilai
Client IP Address	IP address dari host yang melakukan permintaan	123.1.2.3
Identh user	Remote logname , field ini hampir selalu bernilai null (”-”)	-
Authenticated User Name	Nama atau identitas dari pengguna yang sudah dikonfirmasi	j.user
Request Date and Time	Tanggal dan waktu dimana permintaan diterima oleh server	[14/Jan/2001 23:59:09 -0800]
Request URI	Alamat halaman yang diminta (URL)	/path/to/resource ?query
Status Code	Kode respon HTTP server	200, 304, 404
Bytes Sent	Besar data yang ditransfer dari server ke klien (dalam bytes)	14378
Referrer	URI dari sumber (biasanya sebuah website) darimana halaman dirujuk	http://www.google. com/search?q=oracle
User Agent	Informasi mengenai browser dan OS yang dipakai pengguna	Mozilla/4.51 [en] (WinNT; U)
Cookie String	Nama Cookie pasangan nilai dipisahkan oleh semicolon	COOKIE1=value; COOKIE2=value
Client Hostname	DNS hostname dari host yang melakukan permintaan	client-123lp. domain.net
Server IP Address	IP address dari host yang memenuhi permintaan	123.1.2.3
Filename	Filename atau alamat yang diminta	index.html
Request Method	Metode permintaan HTTP	GET, POST
Transport Protocol	Protokol HTTP	HTTP/1.1

Pembersihan data dimaksudkan untuk membersihkan file log server sebagai masukan dari pembuatan model nantinya, dari data yang tidak diperlukan dalam proses prediksi kunjungan pengguna ke sebuah halaman web.

File Log server dibangun oleh web server dalam format/bentuk text file berstandar yang disebut dengan *Common Log Format*(CLF). Karena text file ini terstandar maka file tersebut bisa digunakan untuk berbagai program analisa web.

Web Server berjenis Apache menggunakan dua format log file yang umum yaitu *Common* dan *Combined*. Untuk format Apache *Common* setiap permintaan pengguna web ditampilkan sebagai sebuah baris terpisah pada web log. Masing masing *field* dipisahkan oleh spasi dan ada yang diapit oleh tanda petik dua. Untuk nilai null direpresentasikan dengan tanda *dash*(-). *Apache Common Log Format* mengandung semua parameter web log dasar tetapi tidak termasuk parameter *Referer, Agent, Time to Serve , Domain Name* atau *Cookie String*. Sedangkan *Apache Combined Log Format* mengandung semua *field* yang ada pada *Common Log Format*, dengan penambahan *field Referrer* dan *User Agent*.

Masing masing field dan penjelasannya pada *Combined Log Format* dijelaskan pada tabel 3.1

Sebelum dilakukan proses pembersihan, File Web log yang berbentuk text file tersebut dirubah kedalam format database untuk mempermudah dalam proses berikutnya. Setelah itu baru dilakukan proses pembersihan dari data yang tidak diperlukan.

Data yang tidak diperlukan tersebut meliputi :

1. Data noise : Data noise adalah data halaman web maupun atribut halaman web yang tidak relevan dalam proses prediksi ini yang dianggap bisa menggganggu proses selanjutnya. Data noise dibagi menjadi dua kategori yaitu noise global dan noise lokal.

   Noise global adalah data yang tidak diperlukan yang berlaku secara umum di web. Contoh data yang termasuk noise global adalah halaman web dari situs bayangan, halaman web yang tercatat dua kali (duplikasi halaman), atau halaman web versi sebelumnya.

   Field-field yang dihasilkan dari file log tidak semuanya digunakan. Maka field yang tidak terpakai dihapus, menyisakan 7 field antara lain remote host, request uri, referer, bytes, user agent dan time stamp.

   Record yang mencatat akses menuju halaman halaman cms (content management system) web tersebut adalah yang berikutnya dihapus. Karena halaman halaman cms bukan termasuk halaman yang diakses pengunjung yang diamati dan tidak diperlukan dalam proses analisa sehingga dianggap sebagai noise

global.

Sedangkan yang dimaksud dengan noise lokal adalah data noise yang menyertai sebuah halaman. Data ini juga tidak relevan dalam proses prediksi. contoh noise lokal adalah gambar, multimedia, panduan navigasi, script. Maka Record dengan ekstensi nama file JPEG, GIF, CSS dan seterusnya, yang dapat ditemukan dalam field URL dari setiap record, dapat dihapus dari record log. File-file dengan ekstensi ini tidak menunjukkan halaman yang menarik pengguna, melainkan hanya dokumen tertanam di halaman web, sehingga file file tersebut tidak diperlukan dalam proses identifikasi halaman web yang diakses pengguna.

2. Record dengan kode status pengiriman gagal: kode pengiriman data tercatat pada field status. Kode status HTTP kemudian dipertimbangkan dalam proses selanjutnya untuk membersihkan. Pada langkah ini field status pada setiap record dalam log akses web diperiksa dan record dengan kode status lebih dari 299 atau di bawah 200 dihapus.
3. Data yang dihasilkan oleh Robot : Sebuah Web Robot (WR), juga disebut spider atau bot, adalah perangkat lunak yang secara berkala memindai situs web untuk membaca isi website. Robot web secara otomatis mengikuti semua hyperlink dari halaman web yang sedang dipindai. Search engine seperti Google, menggunakan WRS untuk mengumpulkan semua halaman dari sebuah situs web untuk memperbarui indeks pencarian mereka. Jumlah permintaan dari satu robot web mungkin sama dengan jumlah URL situs Web. Jika situs web tidak menarik banyak pengunjung, maka jumlah dari permintaan yang datang dari semua robot web yang telah mengunjungi situs mungkin melebihi dari permintaan yang dihasilkan oleh manusia. Untuk mengidentifikasi permintaan yang dilakukan oleh bot, bukan pengguna manusia, maka record yang mengandung string *robots.txt* pada field *request uri* dihapus.

3.1.2 Identifikasi Pengguna

Identifikasi pengguna adalah proses mengidentifikasi setiap pengguna yang berbeda yang mengakses situs Web. Tujuan identifikasi pengguna untuk mendapatkan karakteristik akses setiap pengguna yang unik, dan kemudian membuat pengelompokan pengguna. Setiap pengguna memiliki alamat IP yang unik dan masing-masing alamat IP mewakili satu pengguna. Tapi sebenarnya ada tiga kondisi yang perlu diperhatikan:

1. Beberapa pengguna memiliki Alamat IP yang unik.

2. Beberapa pengguna bisa memiliki dua atau lebih alamat IP.
3. Karena adanya server proxy, beberapa pengguna dapat berbagi satu alamat IP

Dari kondisi tersebut untuk identifikasi pengguna dilakukan aturan sebagai berikut :

1. alamat IP yang berbeda merujuk ke pengguna yang berbeda.
2. IP yang sama dengan sistem operasi yang berbeda atau browser yang berbeda harus dipertimbangkan sebagai pengguna yang berbeda.
3. Sedangkan alamat IP, sistem operasi dan browser yang semua sama, pengguna baru dapat ditentukan dari apakah halaman yang diminta dicapai dengan halaman yang diakses sebelumnya.

Sebagai contoh, misalkan dari proses pembersihan diperoleh data log seperti pada Tabel 3.2. Untuk memudahkan dalam pembacaan dan penyimpanan data maka setiap URL dinotasikan ke dalam alphabet.

Tabel 3.2 Contoh Data Log

Time	**User Agent**	**IP**	**URL**	**Referer**
0:01	Mozilla/5.0 (Linux; Android 4.2.2; en-us; SAMSUNG GT-I9500 Build/JDQ39)	1.2.3.4	p1	-
0:09	Mozilla/5.0 (Linux; Android 4.2.2; en-us; SAMSUNG GT-I9500 Build/JDQ39)	1.2.3.4	p2	p1
0:10	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X)	1.2.3.5	p3	p1
0:12	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X)	1.2.3.5	p6	p3
1:15	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p1	-
1:25	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p7	p1
2:30	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p2	p7

Pada proses identifikasi pengguna, dengan melihat IP dan User Agent yang digunakan seperti aturan identifikasi pengguna diatas, maka dihasilkan 3 kelompok pengguna seperti tampak pada Tabel 3.3.

Tabel 3.3 Contoh Identifikasi Pengguna

Pengguna	Time	User Agent	IP	URL	Referer
Pengguna 1	0:01	Mozilla/5.0 (Linux;Android 4.2.2; en-us;SAMSUNG GT-I9500 Build/JDQ39)	1.2.3.4	p1	-
Pengguna 1	0:09	Mozilla/5.0 (Linux;Android 4.2.2; en-us;SAMSUNG GT-I9500 Build/JDQ39)	1.2.3.4	p2	p1
Pengguna 2	0:10	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X)	1.2.3.5	p3	p1
Pengguna 2	0:12	Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X)	1.2.3.5	p6	p3
Pengguna 3	1:15	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p1	-
Pengguna 3	1:25	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p7	p1
Pengguna 3	2:30	Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0	1.2.3.5	p2	p7

Penjelasan pengelompokan ini sebagai berikut :

1. Record 1, 2 merupakan record yang dimiliki oleh pengguna 1 karena record 1 dan record 2 menggunakan IP dan user agent yang sama yaitu IP 1.2.3.4 dan User Agent Mozilla/5.0 (Linux; Android 4.2.2; en-us; SAMSUNG GT-I9500 Build/JDQ39) .
2. Record 3 dan 4 dimiliki oleh pengguna yang lain dari record 1 dan 2, karena IP yang digunakan berbeda dari record sebelumnya. Record 3 dan 4 yang menggunakan IP 1.2.3.5 dan menggunakan user agent Mozilla/5.0 (iPhone; CPU iPhone OS 6_0 like Mac OS X) diidentifikasi digunakan oleh pengguna 2.
3. Record 5,6 dan 7 meskipun menggunakan IP yang sama dengan 2 record se-

belumnya namun karena user agent menunjukkan hasil yang berbeda, maka record 5,6 dan 7 dianggap dihasilkan dari pengguna yang berbeda dari record sebelumnya. Record 5,6,7 dimiliki pengguna 3 yang menggunakan IP 1.2.3.5 dan menggunakan user agent Mozilla/5.0 (Windows NT 6.1; rv:26.0) Gecko/20100101 Firefox/26.0

3.1.3 Identifikasi Sesi

Sesi adalah durasi waktu yang digunakan untuk mengakses halamanhalaman web. Sedangkan sebuah sesi pengguna didefinisikan sebagai urutan permintaan yang dilakukan oleh seorang pengguna selama beberapa periode waktu. [Liu, 2006] Tujuan dari identifikasi sesi ini adalah untuk membagi halaman-halaman yang diakses setiap pengguna menjadi sesi- sesi. Pada metode untuk mengidentifikasi sesi pengguna digunakan juga mekanisme timeout dan referensi kunjungan.

Berikut ini adalah aturan yang digunakan untuk mengidentifikasi sesi pengguna :

1. Jika ada user baru maka dibentuk sesi baru;
2. Dalam satu sesi pengguna, jika perujuk halaman(referer) bernilai null atau jika perujuk halaman bernilai tidak sama dengan halaman yang diakses sebelumnya, maka dibentuk sesi baru;
3. Jika waktu antara permintaan suatu halaman dengan permintaan halaman sebelumnya melebihi batas tertentu (dalam penelitian ini digunakan batas waktu 30 menit), diasumsikan bahwa pengguna memulai sesi baru

Untuk memperjelas pembentukan sesi pengguna akan digunakan ilustrasi contoh sebagai berikut.

Tabel 3.4 Contoh Identifikasi Sesi

Sesi	Time	IP	URL	Referer
Sesi 1	0:01	1.2.3.4	p1	-
Sesi 1	0:09	1.2.3.4	p2	p1
Sesi 2	0:10	1.2.3.4	p3	p1
Sesi 2	0:12	1.2.3.4	p6	p3
Sesi 3	1:15	1.2.3.4	p1	-
Sesi 3	1:25	1.2.3.4	p7	p1
Sesi 4	2:30	1.2.3.4	p2	p7

Dengan menggunakan data pada Tabel 3.3 dilakukan proses identifikasi

sesi yang kemudian menghasilkan 4 sesi pengguna seperti pada Tabel 3.4. Identifikasi dilakukan dengan membandingkan Id Pengguna, Referer dan menghitung selisih waktu akses sebuah record dengan record sebelumnya.

Penjelasan proses identifikasi sesi untuk contoh diatas adalah sebagai berikut :

1. Record 1, karena merupakan pengguna baru maka dibentuk sesi baru dan diberi ID sesi 1.
2. Record 2 karena memiliki ID pengguna yang sama dengan record1, memiliki referer yang sama dengan URL record 1 serta selisih waktu antara time record 1 dengan record 2 kurang dari 30 menit, maka record 2 diberi ID sesi yang sama dengan record 1, yaitu sesi 1.
3. record 3, karena memiliki ID pengguna yang berbeda dengan record 2 maka diberi ID yang berbeda yaitu sesi 2.
4. Record 4 karena memiliki ID pengguna yang sama dengan record1, memiliki referer yang sama dengan URL record 3 serta selisih waktu antara time record 3 dengan time record 4 kurang dari 30 menit, maka record 4 diberi ID sesi yang sama dengan record 3, yaitu sesi 2.
5. record 5, karena memiliki ID pengguna yang berbeda dengan record 4, maka diberi ID sesi yang berbeda, yaitu sesi 3.
6. Record 6 karena memiliki ID pengguna yang sama dengan record 5, memiliki referer yang sama dengan URL record 5 serta selisih waktu antara time record 5 dengan time record 6 kurang dari 30 menit, maka record 6 diberi ID sesi yang sama dengan record 5, yaitu sesi 3.
7. Record 7 meskipun memiliki ID pengguna yang sama dengan record 6, memiliki referer yang sama dengan URL record 6, namun selisih waktu antara time record 7 dengan time record 6 lebih dari 30 menit, maka record 7 diberi ID sesi yang berbeda dengan record 6, yaitu sesi 4.

Setiap sesi pengguna merupakan urutan permintaan yang menggambarkan halaman(URL) yang dikunjungi oleh seorang pengguna. Sehingga setiap sesi pengguna dinotasikan dalam urutan alphabet dengan tanda koma sebagai pemisah setiap URL.

Dari data identifikasi sesi diatas, semua URL yang diakses pada setiap sesi dirubah dalam bentuk urutan. Maka kumpulan sesi pengguna yang diperoleh seperti tampak pada Tabel 3.5

Tabel 3.5 Sesi Pengguna

Sesi	Time	IP	URL
Sesi 1	0:01	1.2.3.4	p1,p2
Sesi 2	0:10	1.2.3.4	p3, p6
Sesi 3	1:15	1.2.3.4	p1,p7
Sesi 4	2:30	1.2.3.4	p2

3.2 Pembuatan n-gram Model

Model n-gram direpresentasikan oleh 3 parameter ¡A,S,T¿, dimana A adalah kumpulan kemungkinan aksi yag mungkin dilakukan oleh pengguna, S adalah sejumlah kemungkinan state dimana model n-gram dibangun dan T adalah nilai kemungkinan pengguna melakukan aksi A ketika proses berada pada state S.

Ruang state dari model n-gram bergantung pada jumlah aksi sebelumnya yang digunakan untuk prediksi aksi berikutnya. Model n-gram yang paling sederhana memprediksi aksi selanjutnya hanya berdasarkan pada aksi terakhir yang dilakukan oleh user. Dalam model ini dikenal dengan sebutan model 1-gram atau model unigram. Model yang lebih komplek melakukan prediksi dengan melihat pada 2 aksi terakhiryang dilakukan oleh pengguna yang disebut model 2-gram atau model bigram dan T sebagai state yang berhubungan pada semua kemungkinan pasangan aksi yang mungkin dilakukan dalam urutan tersebut. Sehingga sebuah n-gram model adalah model yang menggunakan n urutan aksi terakhir untuk melakukan prediksi aksi berikutnya dari pengguna.

Masukan untuk pembuatan model n-gram adalah kumpulan sesi pengguna yang dihasilkan dari pra proses yang sudah dijelaskan pada sub bab 3.1 . Model n-gram berhubungan dengan halaman halaman yang berbeda yang ada pada website dan setiap state merepresentasikan semua urutan halaman dengan panjang n yang sudah diobservasi dari kumpulan sesi pengguna. Setelah state pengguna diidentifikasi, berikutnya akan dihitung jumlah frekuansi kejadian dari masih masing state. Untuk membangun model n-gram ini digunakan training set dari kumpulan sesi pengguna.

Adapun langkah-langkah untuk membangun model n-gram dijelaskan sebagai berikut :

1. Baca record sesi pengguna sebagai L, Model n-gram yang dibangun sebagai H[n];

2. Hitung panjang L sebagai t, jika nilai t lebih besar dari n lanjutkan ke proses berikutnya.
3. Catat state model ke tabel H dengan mengambil L mulai dari substring ke 1 sampai substring ke n sebagai P.
4. Catat substring ke n+1 untuk aksi berikutnya sebagai C
5. Tambahkan nilai 1 untuk frekuensi kejadian P $\rightarrow$ C sebagai F[P,C]
6. Bandingkan nilai F[P,C] dengan nilai kejadian tertinggi dari state P, Max[P], jika nilai F[P,C] lebih besar dari Max[P] maka catat C sebagai aksi berikutnya untuk state P pada tabel H[n].
7. Hentikan proses sampai semua record sesi pengguna terbaca.

Untuk lebih memperjelas langkah-langkah pembentukan model n-gram diatas, akan digunakan penjelasan dengan ilustrasi contoh berikut ini. Misalkan kumpulan sesi pengguna yang dimiliki, seperti pada tabel 3.6

Tabel 3.6 Sesi Pengguna

ID Sesi	**Sesi Pengguna**
S1	P3 , P2 , P1
S2	P3, P5, P2, P1, P4
S3	P4 , P5 , P2 , P1 , P5 , P4
S4	P3 , P4 , P5 , P2 , P1
S5	P1, P4, P2, P5, P4

Sebagai contoh pada sesi pengguna S2 (P3 , P5 , P2 , P1 , P4) dari tabel 3.6. Jika dibangun model 1-gram berdasarkan algoritma diatas, maka setiap state dibuat dari sebuah aksi tunggal, sehingga state P dan aksi berikutnya C yang dihasilkan adalah P3 $\rightarrow$ P5, P5$\rightarrow$ P2, P2$\rightarrow$ P1, dan P1 $\rightarrow$ P4. Sedangkan jika mengambil sesi pengguna S3 (P4 , P5 , P2 , P1 , P5 , P4), maka pasangan P$\rightarrow$C yang dihasilkan adalah , P4$\rightarrow$P5, P5$\rightarrow$P2, P2$\rightarrow$P1, P1$\rightarrow$P5 dan P5$\rightarrow$P4.

Setelah semua state diidentifikasi, berikutnya untuk setiap pasangan P$\rightarrow$C, akan dihitung nilai F untuk pasangan tersebut. Dimana F adalah frekuensi kejadian P $\rightarrow$ C. Jika menggunakan sesi pengguna S3 saja akan didapat model dengan parameter (P$\rightarrow$C:F) antara lain P3 $\rightarrow$ P5:1, P5$\rightarrow$P2:1, P2$\rightarrow$ P1:1, dan P1$\rightarrow$P4:1.

Sedangkan jika menggunakan seluruh data pada tabel 3.6, akan diperoleh data model dengan parameter (P,C,F) selengkapnya seperti yang ditampilkan pada tabel 3.7

Tabel 3.7 Model 1-gram dengan parameter P,C,F

State	Aksi Berikutnya	Frekuensi Kejadian
P1	P4	2
P1	P5	1
P2	P1	4
P2	P5	1
P3	P2	1
P3	P4	1
P3	P5	1
P4	P2	1
P4	P5	2
P5	P2	3
P5	P4	2

Berikutnya untuk setiap State P dibandingkan nilai F yang dimiliki oleh masing masing pasangan P→ C. Nilai F yang tertinggi yang akan digunakan sebagai model dan yang lainnya akan dihapus. Misalkan untuk State P5, nilai F untuk P5→P2 lebih besar, (yaitu 3) jika dibandingkan nilai F untuk P5 → P4 (yaitu 2), maka model yang digunakan adalah P5→ P2 dan P5→P4 akan dihapus. Sehingga dengan menggunakan seluruh data pada Tabel 3.7 akan diperoleh model 1-gram seperti pada Tabel 3.8

Tabel 3.8 Model 1-gram

State	Aksi Berikutnya	Frekuensi Kejadian
P1	P4	2
P2	P1	4
P3	P2	1
P4	P5	2
P5	P2	3

Sedangkan untuk membangun model 2-gram akan diberikan contoh berikut, dengan menggunakan data sesi pengguna yang sama yaitu Tabel 3.6. Untuk sesi pengguna S2 (P3 , P5 , P2 , P1 , P4), maka setiap state dibuat dari 2 buah aksi yang berurutan , sehingga state P dan aksi berikutnya C yang dihasilkan adalah P3,P5 → P2; P5,P2→P1 dan P2,P1→ P4. Untuk pasangan P→ C,F dimana F

adalah frekuensi kejadian P → C, akan diperoleh P3,P5 → P2:1 ; P5,P2→P1:1 dan P2,P1→ P4:1. Sedangkan jika mengambil sesi pengguna S3 (P4 , P5 , P2 , P1 , P5 , P4), maka pasangan P→C yang dihasilkan adalah , P4,P5→P2, P5,P2 → P1, P2,P1 → P1 dan P1,P5 →P4.

Setelah semua state diidentifikasi dan frekuensi kejadian dihitung, dilakukan mekanisme yang sama seperti membangun model 1-gram diatas. Untuk setiap State P dibandingkan nilai F yang dimiliki oleh masing masing pasangan P→ C. State P dengan nilai F yang tertinggi yang akan digunakan sebagai model dan state P yang sama yang memiliki nilai lebih rendah akan dihapus dari model.

Sehingga dengan menggunakan seluruh data akan diperoleh model 2-gram seperti pada Tabel 3.9

Tabel 3.9 Model 2-gram

State	Aksi Berikutnya	Frekuensi Kejadian
P1,P5	P4	1
P1,P4	P2	1
P2,P1	P4	1
P2,P5	P4	1
P3,P2	P1	1
P3,P5	P2	1
P3,P4	P5	1
P4,P5	P2	2
P4,P2	P5	1
P5,P2	P1	3

3.3 Pruning Model

Untuk mengurangi besar model yang dihasilkan digunakan skema pruning yang biasanya dipakai pada metode decision tree. Jadi tujuan dari proses penghapusan (pruning) ini adalah untuk mengurangi kompleksitas model tetapi bisa mempertahankan bahkan menaikkan tingkat akurasi prediksi dan juga mempertahankan aplicability model. Teknik ini didasarkan pada pengamatan bahwa untuk sebuah state data uji yang diprediksi aksi berikutnya, ada beberapa state yang bisa digunakan dari berbagai n-gram model. Masing masing model n-gram bisa memberikan prediksi yang berbeda yang juga menghasilkan akurasi yang berbeda pula. Maka penentuan model mana yang digunakan dalam prediksi akan sangat menen-

tukan tingkat akurasi. Skema yang diambil adalah menghapus state pada model yang kemungkinan memiliki akurasi prediksi yang rendah. Dengan menghilangkan state-state tersebut akan mengurangi kompleksitas model. State-state yang lolos dari proses penghapusan (pruning) yang nantinya akan digunakan dalam proses prediksi. Ada 2 skema yang digunakan dalam proses ini, yaitu :

1. Support Pruning : skema ini menghapus state berdasarkan nilai support yang memiliki masing masing state.
2. Error Pruning : Skema ini menggunakan pendekatan penghapusan berbasis error untuk menghapus state dengan akurasi prediksi rendah.

Dalam penelitian ini akan dicoba menggunakan 2 skema tersebut dan akan dibandingkan hasilnya. Penjelasan mengenai kedua skema tersebut dijelaskan sebagai berikut :

3.3.1 Support Pruning

Support pruning didasarkan pada pengamatan bahwa string string yang memiliki dukungan yang rendah pada training set cenderung juga memiliki akurasi prediksi yang rendah. Sehinga string yang memiliki dukungan(support) rendah bisa dihilangkan tanpa mempengaruhi akurasi keseluruhan dan cakupan model yang dihasilkan. Support disini adalah jumlah kejadian yang dimiliki setiap state. Jumlah pemangkasan dalam skema support pruning ini dikendalikan oleh parameter yang disebut sebagai frekuensi threshold. Ada beberapa kemungkinan tentang skema support pruning ini. Pertama, frekuensi threshold yang sama diterapkan pada state tanpa memperhatikan tingkat oredernya. Skema yang kedua, kebijakan pemangkasan ini hanya diberlakukan pada string di order tinggi dengan asumsi atring yang lebih tinggi memiliki nilai dukungan yang rendah sehingga dapat mengurangi kompleksitas model. Pada penelitian ini akan digunakan skema pertama dimana semua gram model akan diterapkan nilai ambang frequnesi yang sama. Untuk setiap state pada semua model n-gram yang memiliki nilai jumlah kejadian dibawah frequensi threshold yang sudah ditetapkan akan dihapus dari model.

3.3.2 Error Pruning

Dalam skema error pruning ini, diukur tingkat error masing masing state untuk memutuskan perlu tidaknya state tersebut dihapus. Untuk memperkirakan error dari masing masing string dilakukan proses validasi. Selama tahap validasi, seluruh model diuji dengan menggunakan bagian dari training set yang disebut set validasi. Set validasi ini tidak digunakan dalam proses pembentukan model. Se-

hingga sebelum proses dilakukan, data sesi pengguna dibagi menjadi 3, 60 % tetap akan digunakan sebagai data training untuk membuat model dan 20% digunakan sebagai set validasi dan 20% digunakan sebagai data uji. Langkah-langkah yang dilakukan pada error pruning dibuat sebagai berikut:

1. Dengan menggunakan data sesi pengguna pada validasi set dibuat model n-gram seperti dijelaskan pada sub bab 3.2. Dan kita sebut sebagai model n-gram validasi
2. Dengan menggunakan data sesi pengguna pada tarining set dibuat model n-gram seperti dijelaskan pada sub bab 3.2. Dan kita sebut sebagai model n-gram training
3. Menghitung prosentase nilai error untuk masing state P→ C pada n-gram model training. Nilai error (P→C) pada model training dihitung dari frekuensi state P pada data validasi dikurangi frekuensi P→ C dibagi dengan frekuensi sate P dikali 100%

$$Error(P \rightarrow C) = \frac{F(P) - F(P \rightarrow C)}{F(P)} x100\% \tag{3.1}$$

4. Selanjutnya, untuk setiap state pada n-gram model tertinggi kami mengidentifikasi sub set yang sama pada orde dibawahnya. Sebagai contoh jika state pada n-gram adalah p5, p3, p6,p7, state pada gram lebih rendah yang diidentifikasi adalah p3,p6,p7 (3-gram), p6,pa7 (2-gram) dan p7 (1-gram). Jika nilai error state p5, p3, p6, p7 pada model 4-gram lebih besar dibandingkan p3, p6, p7 (3-gram) atau p6,pa7 (2-gram) atau p7 (1-gram) maka state tersebut dihapus.
5. Lakukan prosedur yang sama untuk mengindentifikasi semua state yang berada pada n-gram yang lebih rendah kecuali state pada model 1-gram. State pada model 1-gram tidak dihapus supaya tidak mengurangi cakupan model secara keseluruhan dalam membuat prediksi.

3.4 Sistem Prediksi

Pada algoritma n-gram tradisonal, prediksi yang dilakukan didasarkan pada satu nilai n saja, bisa menggunakan model 1-gram, 2-gram atau 3-gram saja dan seterusnya. Misalkan digunakan model 2-gram maka akan diambil dua urutan aksi terakhir aksi oleh pengguna kemudian mencari urutan aksi yang sama dengan state pada model, prediksi diberikan berdasarkan aksi berikutnya yang tercatat dari state tersebut pada model.

Dalam penelitian ini akan digunakan skema n-gram+ seperti yang digunakan oleh [Zhong Su, 2000] .Skema ini didasarkan pada beberapa penelitian yang menyebutkan, model 1-gram atau biasa disebut model unigram tidak menghasilkan prediksi kunjungan halaman berikutnya dengan hasil memuaskan. Hal ini karena model unigram tidak melihat terlalu jauh catatan masa lalu pengguna untuk benar benar bisa membedakan perilaku pengguna. Sehingga untuk mengatasi hal tersebut digunakan model dengan berbagai tingkatan gram yang lebih tinggi, misalnya model 3-gram, 4-gram dan seterusnya. Namun, ada beberapa data set yang tidak memiliki jumlah string yang cukup banyak pada n-gram order tinggi yang mengakibatkan cakupan yang tidak luas karena sedikitnya jumlah model yang dihasilkan sehingga juga menurunkan tingkat akurasi prediksi yang dihasilkan dari model model dengan gram yang tinggi.

Dalam skema n-gram+ , untuk setiap data uji , model n-gram yang mencakup contoh data uji digunakan terlebih dahulu. Sebagai contoh misal dibangun 3 model yaitu model 3-gram, model 2-gram dan model 1-gram. Ketika diberikan sebuah data uji, jika panjang state data uji diatas 3 untuk pertama kali akan dibuat prediksi dengan menggunakan model 3-gram. Namun jika model 3-gram tidak memiliki state yang sesuai atau panjang state data uji kurang dari 3, maka dicoba membuat prediksi dengan menggunakan model 2-gram, dan seterusnya. Urutan langkah tersebut seperti digambarkan pada Gambar *flowchart* 3.3.

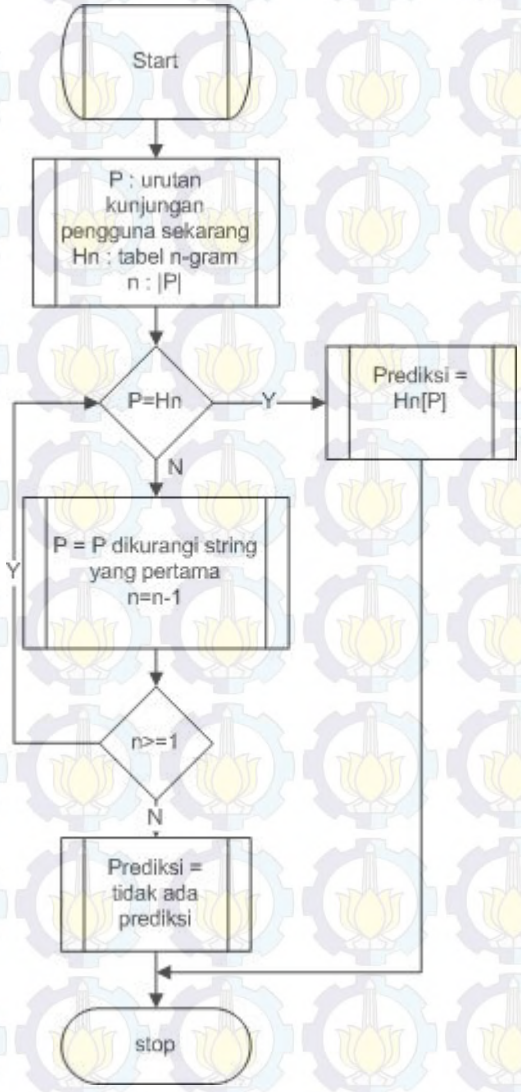
Penjelasan untuk urutan langkah untuk membuat prediksi dilakukan adalah sebagai berikut :

1. Baca record sesi pengguna yang diuji sebagai L,
2. Potong state pada L dengan mengambil 3 substring terakhir.
3. Bandingkan L dengan state yang sama pada model 3-gram, jika state ditemukan berikan prediksi dari state tersebut
4. Jika tidak ditemukan state yang sesuai, ambil 2 substring terakhir dari L.
5. Ulangi langkah 3 dengan menggunakan model gram dibawahnya yaitu model 2-gram
6. Jika pada model 1-gram state tidak ditemukan berikan nilai model tidak memberikan prediksi

Pada penelitian ini juga akan dilakukan prediksi dengan algoritma n-gram tradisional sebagai bahan perbandingan. Urutan langkah untuk prediksi dengan n-gram tradisional adalah sebagai berikut :

1. Baca record sesi pengguna yang diuji sebagai L,
2. Potong state pada L dengan mengambil n substring terakhir sesuai dengan model n-gram yang digunakan.

3. Bandingkan L dengan state yang sama pada sebuah model n-gram, jika state ditemukan berikan prediksi dari state tersebut
4. Jika tidak ditemukan state yang sesuai, berikan nilai model tidak memberikan prediksi



Gambar 3.3 Tahapan Praproses

3.5 Evaluasi

Untuk melakukan ujicoba data set yang dihasilkan dari praproses dibagi menjadi training set dan testing set. Untuk skema mengguna error pruning data set dibagi menjadi 3 yaitu, training set, testing set dan validasi set. Prosedur pengujian prediksi dilakukan dengan memotong substring terakhir pada sesi pengguna dari

testing set. Substring ini kemudian dibandingkan dengan hasil prediksi model untuk state yang sama.

Dalam beberapa kasus, model n-gram mungkin tidak dapat membuat prediksi untuk sesi di tes set. Hal ini bisa terjadi karena ada dua kemungkinan, yaitu panjang sesi kurang dari panjang n-gram model, atau sesi tersebut tidak ditemukan dalam training set. Dalam kasus seperti ini, maka akan disebutkan model tidak membuat prediksi. Sehingga akan ada tiga nilai yang dikeluarkan dalam prediksi, yaitu prediksi salah, prediksi benar dan model tidak membuat prediksi.

Ada tiga parameter yang digunakan untuk mengukur efektifitas kinerja model n-gram pada penelitian ini, yaitu :

1. Akurasi (ketepatan prediksi). Secara umum akurasi dari model ini adalah jumlah prediksi yang benar dibagi dengan jumlah prediksi yang dibuat.[Kurian, 2008]

$$Akurasi = \frac{P^+}{P^+ + P^-} \tag{3.2}$$

   Dimana :
   P^+ : prediksi yang betul
   P^- : prediksi yang salah
2. Besar/kompleksitas model. Kompleksitas model dihitung dari jumlah state dari model n-gram yang digunakan dalam membuat prediksi.
3. Aplicability (Cakupan model): mengukur berapa kali model mampu membuat prediksi. [Zhong Su, 2000]

$$Aplicability = \frac{P^+ + P^-}{|R|} \tag{3.3}$$

   Dimana :
   P^+ : prediksi yang betul
   P^- : prediksi yang salah
   $|R|$: total data yang diuji

[*Halaman ini sengaja dikosongkan.*]

BAB 4

Analisa Hasil dan Evaluasi

4.1 Pengambilan Data

Log File yang digunakan dalam penelitian ini diambil dari web server berjenis apache dari sebuah website pemerintah selama bulan januari sampai bulan maret 2014. Untuk memudahkan pengolahan data log file yang berformat text file tersebut dirubah ke dalam tabel database dengan menggunakan php dan database MySql.

Field-field pada tabel antara lain remote host,ident user, auth user, request method, request protocol, request uri (URL), referrer, bytes, status, user agent, time stamp. Contoh data log yang dihasilkan seperti tampak pada Tabel 4.1.

Dari pemindahan file log server selama 3 bulan ke tabel database, diperoleh total data sebanyak 2.876.579 record , dengan perincian record bulan januari sebanyak 1.009.805, record bulan pebruari sebanyak 773.418 dan record bulan maret sebanyak 1.093.356. Data log selama 3 bulan dijadikan satu untuk berikutnya akan diolah dengan proses seperti yang dijelas pada bab 3. Data tidak dipisahkan per bulan karena dalam penelitian ini tidak memperhatikan tren data per bulan, hanya diambil bentuk urutan kunjungan pengguna website yang diuji secara keseluruhan.

4.2 Pra Proses

Pembersihan Data

Tahap berikutnya adalah membersihkan tabel data log dari data-data yang tidak dipakai dalam proses pembuatan model n-gram dan proses prediksi.

Dalam pembuatan model prediksi ini field field yang dipakai hanya 7 field seperti dijelaskan pada sub bab 3.1.1. Maka fieldfield selain 7 file tersebut akan dihapus. Proses berikutnya adalah membersihkan data yang tidak terpakai. Dalam membuat sistem prediksi ini yang dikaji adalah halaman yang dikunjungi oleh pengguna (manusia). Maka yang dimaksud data yang tidak terpakai adalah record yang tidak menunjukkan halaman yang dikunjungi oleh seorang pengguna(manusia).

Kelompok pertama dari record yang dihapus adalah file gambar. Record yang pada *field request uri* mengandung kata jpg , gif, ico, jpeg, png atau ekstensi file multimedia lain dihapus. Ditemukan sekitar 939.986 record file berjenis

Tabel 4.1 File Log Server

remote_host	time_stamp	request_uri	user_agent	referer	status	bytes
82.145.217.209	31/01/2014 13:15	/site/objek-wisata/kabupaten-sidoarjo/	Opera/9.80 (Android; Opera Mini/6.5.27452/34.1088; U; en) Presto/2.8.119 Version/11.10	http://www.google.com/m?q=wisata+daerah+sidoarjo	200	26313
125.167.122.213	31/01/2014 13:20	/site/lambang-daerah/	Mozilla/5.0 (Windows NT 6.2) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/32.0.1700.102 Safari/537.36	https://www.google.com/	200	24959
180.248.56.255	31/01/2014 13:24	/site/gubernur-minta-arkeolog-kembangkan-peninggalan-majapahit/	Mozilla/5.0 (Windows NT 5.1; rv:8.0.1) Gecko/20100101 Firefox/8.0.1	http://www.google.com/url?sa=t&rct=j&q=&esrc=s&sourc	200	29342
180.248.56.255	13/12/2014 15:43	/site/beras-dan-jagung-jatim-tetap-surplus/	Mozilla/5.0 (Windows NT 5.1; rv:8.0.1) Gecko/20100101 Firefox/8.0.1	/site/gubernur-minta-arkeolog-kembangkan-peninggalan-majapahit/	200	29721

gambar/multimedia

Kelompok kedua adalah record yang mengandung string js atau css yang merupakan file script, yang akan dihapus. Ditemukan sekitar 320.282 record file berjenis script.

Kelompok ketiga adalah file halaman cms *(content management system)*. Ditemukan 1.574.429 record halaman cms dari data

Kelompok keempat adalah record yang mencatat akses yang dilakukan bukan pengguna, seperti web robot atau bot maka request uri yang mengandung string robots.txt dihapus. Ditemukan sebanyak 2,869 record dari permintaan yang dilakukan oleh bot.

Kelompok kelima yang dihapus adalah record yang pengirimannya data-nya gagal, maka record yang pada field status-nya bernilai kurang dari 200 dan lebih dari 299 yang berikutnya harus dihilangkan. Ditemukan 250,139 record data dengan status dibawah 200 atau status diatas 299

Setelah proses pembersihan data dilakukan, jumlah data total yang tersisa sebanyak 247.055 record, dengan jumlah IP unik yang mengakses sebanyak 41.834 dan jumlah URL unik yang diakses sebanyak 1.613.

4.2.1 Membangun Sesi Pengguna

Dari data log file yang sudah dibersihkan, dibangun sesi pengguna dengan skema identifikasi pengguna dan identifikasi sesi seperti yang sudah dijelaskan pada sub bab 3.1.2 dan sub 3.1.3

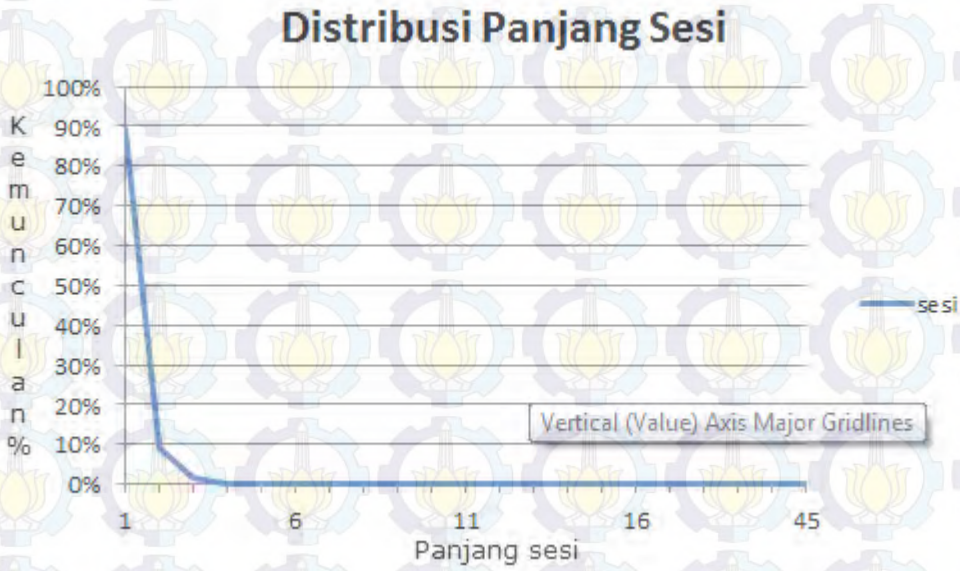
Tabel 4.2 Data sesi pengguna berdasarkan panjang state

Panjang state	Jumlah state
1	190.156
2	19.352
3	2.901
4	452
5	150
6	66
7	33
8	19
>8	52

Hasil akhir dari proses membangun sesi pengguna ini diperoleh total se-

banyak 213.181 sesi pengguna. Ketika seluruh record dikelompokkan berdasarkan panjang urutan state yang dimiliki didapatkan data sesi pengguna berdasarkan panjang state seperti pada Tabel 4.2.

Berdasarkan data pada Tabel 4.2, jumlah state terbanyak adalah state dengan panjang 1 mencapai 90% dari total data. Untuk lebih memudahkan pembacaan data, data divisualisasikan dalam bentuk grafik seperti tampak pada Gambar 4.2 Distribusi panjang sesi pengguna.



Gambar 4.1 Distribusi Panjang Sesi Pengguna

Dari Gambar 4.2 dapat dilihat bahwa jumlah state semakin sedikit untuk panjang state yang semakin banyak dengan penurunan jumlah paling tajam terjadi dari state dengan panjang 1 ke state panjang 2 dan 3. Untuk state dengan panjang 4 keatas terjadi penurunan nanun tidak terlalu banyak. Sehingga bisa ditarik kesimpulan bahwa sebagian besar urutan aksi yang dilakukan oleh pengguna pada website yang diuji berbentuk urutan pendek dengan jumlah urutan terbanyak adalah 1 diikuti jumlah urutan 2 dan 3.

Dari data sesi pengguna yang diperoleh, 20% data diambil secara acak untuk dijadikan sebagai data uji dan 80% sisanya dijadikan sebagai data training. Pada masing - masing kelompok data,record dengan panjang state 1 dihapus dari tabel, karena untuk membuat n-gram model, panjang minimal state yang diperlukan adalah 2.

Setelah proses penghapusan, untuk data training diperoleh 18.420 sesi pengguna, dan data uji memiliki 4.605 sesi pengguna.

Ketika dikelompokkan berdasarkan panjang state pada data training diperoleh data seperti pada Tabel 4.3 :

Tabel 4.3 sesi pengguna berdasarkan panjang state pada data training

Panjang state	Jumlah state
2	12.191
3	1.906
4	268
5	88
6	37
7	25
8	12
9	10
>9	23

4.3 Pembuatan n-gram Model

Dari 18.420 sesi pengguna pada data training dibangun n-gram model dengan menggunakan algoritma seperti yang dijelaskan pada sub bab 3.2

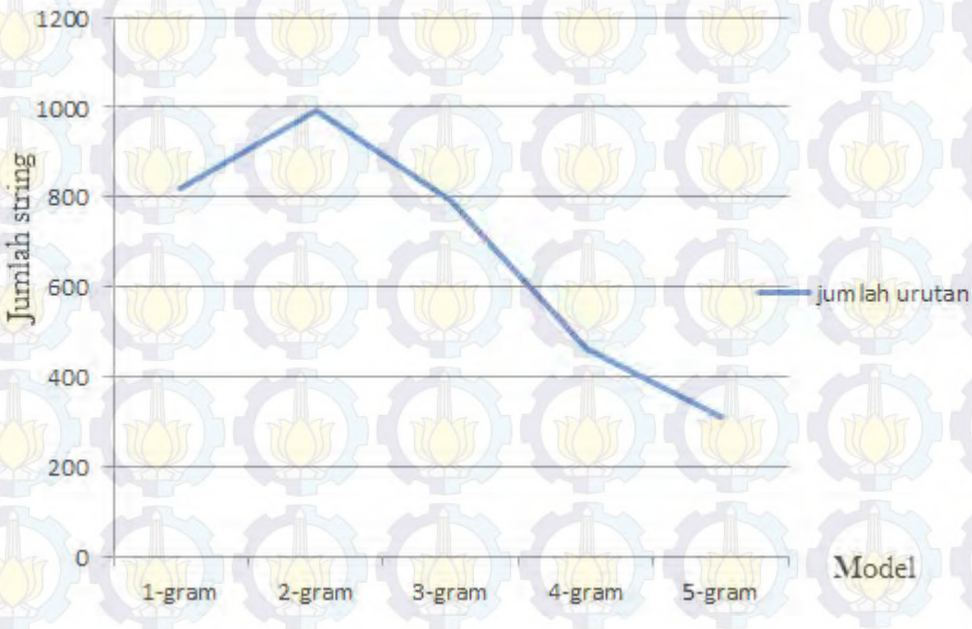
Hasil akhir dari proses ini dihasilkan total 5.974 state pada berbagai tingkat gram model. Adapun besar model(jumlah state) untuk masing-masing model seperti pada Tabel 4.4

Tabel 4.4 Jumlah model n-gram

Model	Jumlah State
1 gram	904
2 gram	1184
3 gram	948
4 gram	573
5 gram	383
>5 gram	1982

Jumlah model terbanyak ada pada model 2-gram sedangkan untuk model 1-gram dan 3-gram jumlah model sedikit dibawah model 2-gram. Jika melihat data

ketika direpresentasikan dalam bentuk grafik seperti pada Gambar 4.3 distribusi n-gram, akan semakin terlihat besar model menurun seiring dengan meningkatnya tingkat gram model mulai pada model 4-gram. Hal ini sudah bisa diprediksikan sebelumnya jika melihat data sesi pengguna pada Tabel 4.6, dimana jumlah state dengan panjang sesi diatas 3 mulai menurun tajam.



Gambar 4.2 Distribusi n-gram

Untuk model n-gram+, state pada masing - masing model n-gram+ diperoleh dari model n-gram ditambah dengan state pada model n-1,n-2...1-gram. Pada penelitian ini kami menggunakan model 1-gram+, 2-gram+, 3-gram+ dan model 4-gram+ untuk analisa kinerja model dalam membuat prediksi.

Tabel 4.5 Jumlah model n-gram+

Model	Jumlah State
1-gram+	904
2-gram+	2088
3-gram+	3036
4-gram+	3609

Jumlah masing - masing state pada model n-gram+ seperti tampak pada Tabel 4.5. Karena model yang digunakan oleh n-gram+ menggunakan semua state

pada model tersebut ditambah dengan semua state pada model gram dibawahnya, maka bisa dilihat bahwa semakin tinggi orde n-gram maka semakin besar jumlah model yang artinya model semakin kompleks.

4.3.1 Proses Pruning

Selain model n-gram dan model n-gram+, dalam penelitian ini akan dilakukan proses pruning pada model 4-gram+.

Tabel 4.6 Besar Model Hasil support pruning

Frekuensi Threshold	Jumlah state model
0	3609
2	926
4	547
6	410
8	334
10	270
12	225
14	199
16	176
18	148
20	139

Model yang dihasilkan dari proses pruning baik dengan support pruning maupun dengan error pruning nantinya akan diuji kinerjanya dalam membuat prediksi dan dibandingkan dengan model n-gram dan model n-gram+ baik dari segi akurasi maupun aplicabilitynya.

Setelah dilakukan skema support pruning pada model 4-gram+ seperti pada penjelasan sub bab 3.3.1, dengan menggunakan ambang frekuensi 0 sampai dengan 20 diperoleh jumlah state pada model 4-gram+ seperti pada Tabel 4.6

Dari penerapan skema support pruning dengan nilai frekuensi threshold 2 yang diterapkan pada model 4-gram+ terlihat sudah mampu mereduksi jumlah state sangat signifikan, dari 3.609 state menjadi hanya 926 state, artinya terjadi penurunan besar mondel mencapai 75%. Pada penerapan frekuensi threshold 4 dan 6 terjadi penurunan besar model rata-rata sebesar 50%. Penurunan besar model tidak terlalu tajam mulai pada penggunaan ambang frekuensi diatas 6.

Model berikutnya dibuat dengan menerapkan skema error pruning terhadap model 4-gram+. Dari proses error pruning ini besar model yang dihasilkan mengalami penurunan sebesar 54% menjadi sebanyak 1.659 state.

4.4 Evaluasi Sistem Prediksi

Setelah model dibuat dan diukur besarnya pada proses diatas, pada tahap evaluasi ini akan diukur kinerja model berdasarkan nilai akurasi dan aplicabilitynya dalam membuat prediksi. Model yang akan diuji meliputi model 1-gram+, 2-gram+, 3-gram+ dan 4-gram+. Sebagai bahan pembanding akan dilakukan juga prediksi dengan menggunakan model 1-gram, 2-gram, 3-gram dan 4-gram. Selain itu terakhir akan diukur juga kinerja dari model4+ yang sudah mengalami tahap pruning baik support pruning maupun error pruning.

Dari proses prediksi dengan menggunakan algoritma n-gram diperoleh nilai akurasi dan aplicability model seperti pada Tabel 4.7

Tabel 4.7 Hasil ujicoba beberapa model n-gram

Model	Akurasi(%)	Aplicability(%)	jumlah state
1 gram	60,28	99,07	904
2 gram	51,92	12,42	1184
3 gram	34,14	0,89	948
4 gram	9,09	0,24	573

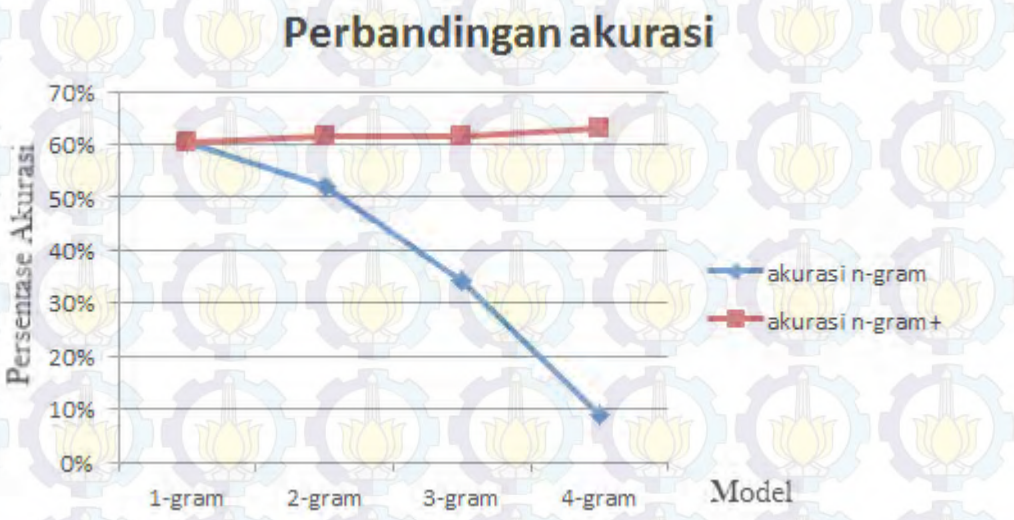
Untuk algoritma n-gram tradisional, kinerja terbaik dihasilkan oleh model 1-gram, baik dari nilai akurasi maupun aplicability. Model 2-gram yang memiliki jumlah state lebih besar justru menghasilkan kinerja lebih buruk baik dari akurasi maupun aplicability model dibandingkan dengan model 1-gram. Begitu juga dengan model 3-gram, memberikan kinerja lebih buruk dari model 1-gram, meskipun jumlah state model 3-gram hampir sama dengan model 1-gram. Jika melihat pada model 4-gram penurunan kinerja ini, kemungkinan disebabkan oleh terlalu sedikitnya jumlah model yang dimiliki model 4-gram untuk memprediksi.

Sedangkan pengujian dengan algoritma n-gram+ diperoleh hasil seperti pada Tabel 4.8

Tabel 4.8 Hasil ujicoba model n-gram+

Model	Akurasi(%)	Aplicability(%)	jumlah state
1-gram+	60,28	99,07	904
2-gram+	61,57	99,06	2.088
3-gram+	61,62	99,07	3.036
4-gram+	62,88	99,78	3.992

Dari hasil pengujian pada model n-gram+ yang ditampilkan pada Tabel 4.8, terlihat bahwa nilai akurasi dan aplicability model dalam membuat prediksi semakin naik seiring semakin naiknya gram model. Dari model 1-gram+, 2-gram+, 3-gram+ dan 4-gram+, kinerja terbaik untuk akurasi dan aplicability model dihasilkan oleh model 4-gram+.

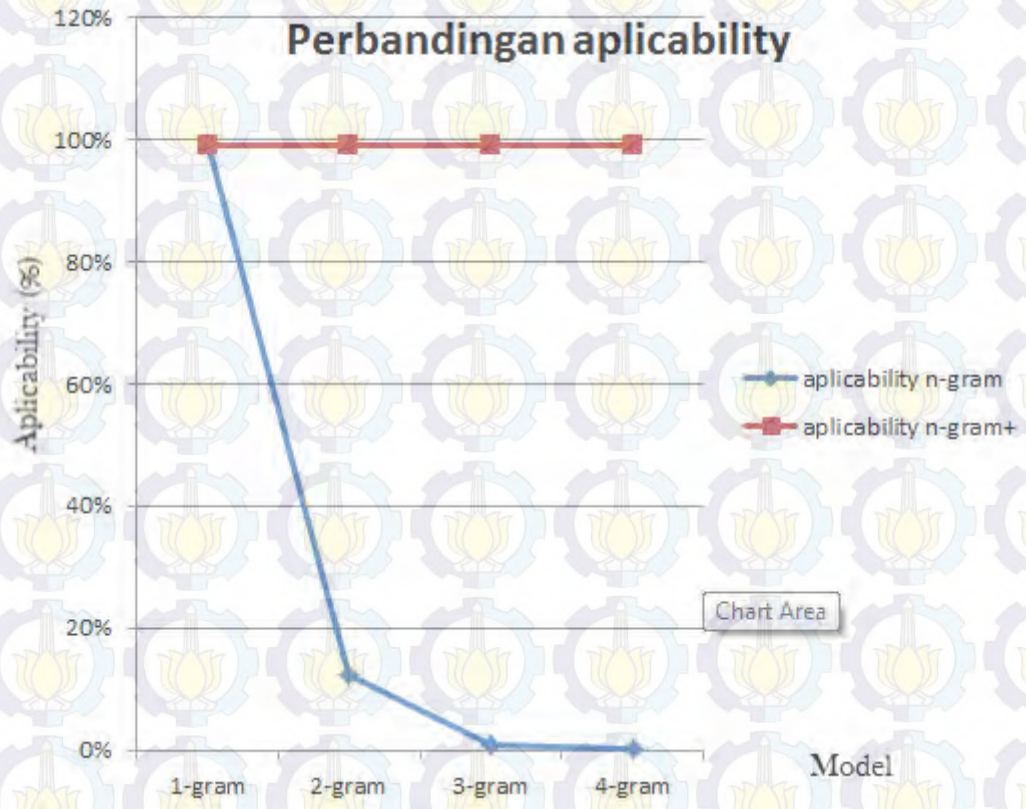


Gambar 4.3 Akurasi n-gram dan n-gram+ model

Untuk melihat lebih jauh kinerja pada model, maka algoritma n-gram tradisional dan algoritma n-gram+ dibandingkan. Untuk perbandingan nilai akurasinya akan tampak seperti pada Gambar 4.3. Secara keseluruhan model n-gram+ menghasilkan akurasi yang lebih baik dibandingkan model n-gram tradisional. Sebagai contoh model 4-gram+ (yang merupakan model terbaik paada algoritma n-gram+) memiliki nilai akurasi 2,60 % lebih baik jika dibandingkan dengan model 1-gram, yang merupakan model yang memberikan hasil terbaik pada algoritma n-gram tradisional. Sehingga bisa ditarik kesimpulan penggunaan algoritma n-gram+ bisa

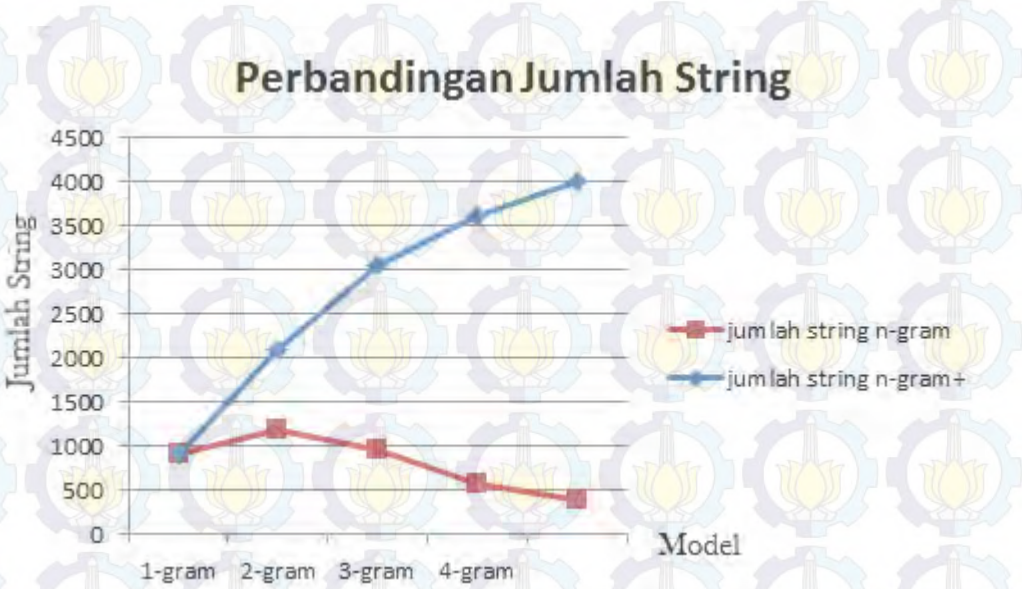
digunakan untuk memperbaiki nilai akurasi model. Semakin tinggi n-gram model semakin besar akurasi yang dihasilkan.

Untuk perbandingan kinerja aplicability model,antara algoritma n-gram dan n-gram+ tampak seperti pada Gambar 4.4. Dengan mengambil contoh yang sama pada pengukuran akurasi, pada model 4-gram+ mampu menaikkan nilai aplicability menjadi 99.78 % jika dibandingkan dengan model 1-gram. Sehingga bisa dilihat bahwa penggunaan algoritma n-gram+ juga bisa menaikkan nilai aplicability model dengan menaikkan jumlah n model.



Gambar 4.4 Aplicability n-gram+ dan n-gram model

Sedangkan terkait masalah besar model, perbandingan algoritma n-gram dan n-gram+ tampak seperti pada Gambar 4.5. Secara keseluruhan, model n-gram+ memberikan hasil kompleksitas model lebih buruk dari model n-gram, karena memang seperti penjelasan diatas bahwa jumlah state pada model n-gram+ didapat dari model n-gram dibawahnya. Sebagai contoh untuk jumlah model pada 4-gram+ mencapai hampir 400% lebih banyak dari model 1-gram.



Gambar 4.5 Jumlah string n-gram+ dan n-gram model

Berikutnya dilakukan pengujian kinerja pada model 4-gram+ yang sudah mengalami proses support pruning.

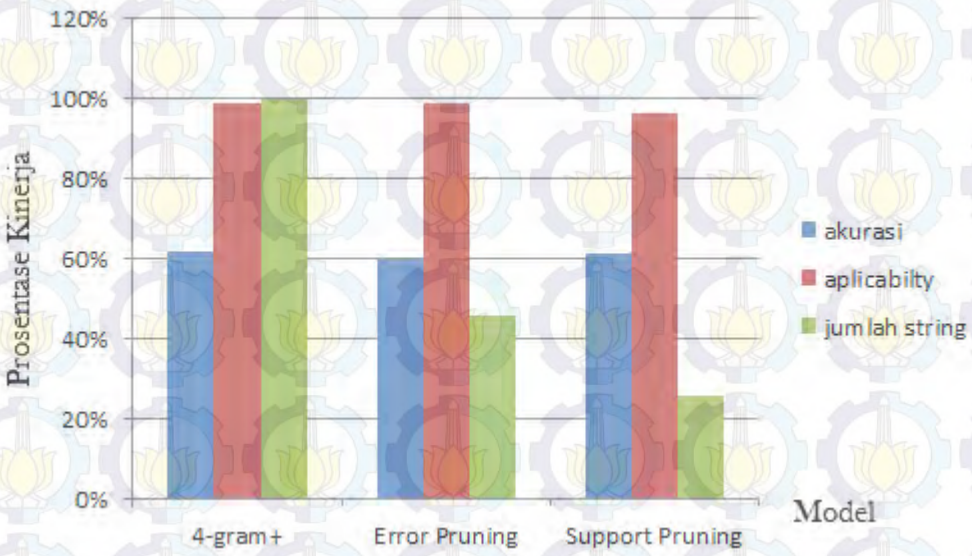
Tabel 4.9 Kinerja model 4-gram+ dengan support pruning

Frekuensi Threshold	Akurasi	Aplicability	Jumlah string
0	62,88	99,78	3609
2	62,36	99,12	826
4	61,17	96,29	547
6	61,14	94,49	410
8	61,01	91,00	334
10	60,68	89,58	270
12	60,30	87,90	225
14	60,15	86,80	199
16	59,92	85,82	176
18	59,64	84,21	148
20	59,52	83,58	139

Hasil pengujian kinerja prediksi yang dihasilkan seperti tampak pada Tabel 4.9. Dari hasil ujicoba untuk penggunaan ambang frequensi 2, nilai akurasi model yang dihasilkan sebesar 62,36%, artinya terjadi penurunan kinerja akurasi sebanyak 0,52% jika dibandingkan dengan akurasi model 4-gram+. Untuk parameter kinerja

aplicability model yang dihasilkan sebesar 99,12%, artinya terjadi penurunan sebesar 0,66%. sedangkan dari segi kompleksitas model terjadi perbaikan dengan menurunnya besar model mencapai 75%. Secara keseluruhan jika dibandingkan dengan model 4-gram+, nilai akurasi dan aplicability model serta besar model mengalami penurunan dengan semakin naiknya ambang frekuensi yang digunakan.

Pada pengujian dengan menggunakan skema error pruning diperoleh hasil nilai akurasi sebesar 60,06% dan aplicability sebesar 98,78%. Jumlah string yang dihasilkan mengalami penurunan menjadi 1.659 string. Dibandingkan dengan kinerja model 4-gram+, skema error pruning menurunkan akurasi model sebesar 2,82%, dan menurunkan aplicability model sebesar 1%. Terjadi perbaikan dari segi kompleksitas model yang menurun mencapai 54%. Hasil kinerja dari error pruning tidak memberikan hasil lebih baik dari pada skema support pruning kemungkinan dikarenakan skema error pruning tidak menyentuh model 1-gram, dimana model 1-gram justru memberikan kontribusi terbesar dalam mebuat prediksi pada kasus data yang memiliki jumlah urutan aksi pendek lebih banyak dibandingkan dengan urutan aksi yang panjang.



Gambar 4.6 Perbandingan kinerja 3 model

Dengan mengambil hasil pengujian model 1-gram sebagai model terbaik pada algoritma n-gram tradisional dan model 4-gram+ yang menghasilkan kinerja terbaik pada algoritma n-gram+ serta model dengan skema pruning akan tampak pada Gambar 4.6 perbandingan kinerja 4 model.

Jika dibandingkan antara model 1-gram dengan hasil model dengan skema

pruning, untuk nilai akurasi model dengan support pruning memiliki nilai akurasi 2% lebih baik, nilai aplicability hampir sama dan kompleksitas model 10% lebih kecil. Maka bisa dilihat bahwa algoritma n-gram+ ditambah dengan skema support pruning bisa menghasilkan model yang tidak kompleks namun mampu mempertahankan akurasi dan aplicabilitynya.

[*Halaman ini sengaja dikosongkan.*]

BAB 5

Kesimpulan dan Pekerjaan Selanjutnya

5.1 Kesimpulan

Berdasarkan hasil ujicoba dan analisis yang telah dilakukan dapat dtarik kesimpulan, untuk sesi pengguna yang memiliki urutan aksi pendek lebih banyak daripada urutan aksi yang panjang, model 1-gram menghasilkan kinerja lebih baik dari model gram lebih tinggi.

Secara keseluruhan model n-gram+ memberikan hasil akurasi dan aplicability lebih baik dari model n-gram tradisional, namun terjadi penurunan kinerja dari segi kompleksitas model. Peningkatan akurasi mencapai 2,6% dan mempertahankan aplicability model diatas 99%, kompleksitas model naik mencapai 400% pada penggunaan model 4-gram+.

Penggunaan support pruning memberikan hasil lebih baik dibandingkan penggunaan error pruning, karena support pruning mampu mengurangi kompleksitas model mencapai 75% namun bisa mempertahankan nilai akurasi dan aplicability model.

Model n-gram+ dengan menggunakan support pruning dari hasil ujicoba secara keseluruhan mampu memperbaiki kompleksitas model namun tetap bisa mempertahankan akurasi dan aplicability model dalam membuat prediksi dibandingkan dengan model n-gram tradisional.

5.2 Pekerjaan Selanjutnya

Pekerjaan selanjutnya yang bisa dilakukan dari penelitian ini adalah :

1. Algoritma n-gram+ perlu diuji coba dengan data set yang berbeda untuk melihat kinerja model, misalkan pada data set yang memiliki urutan aksi yang panjang dalam jumlah yang cukup banyak.
2. membuat sistem prediksi dengan algoritma model n-gram+ dengan menghitung waktu komputasi model dalam memberikan prediksi dan penggunaan memori.
3. menerapkan model n-gram pada algoritma sistem cache.

[*Halaman ini sengaja dikosongkan.*]

Daftar Pustaka

Ajeetkumar S. Patel, A. P. K. (2014). Genetic algorithm and statistical markov model fusion for predicting users surfing behavior using sequence patten mining. *International Journal of Emerging Trends and Technology in Computer Science*, 3.

Chuibi Huang, Jinlin Wang, H. D. J. C. (2013). Mining web logs with plsa based prediction model to improve web caching performance. *JOURNAL OF COMPUTERS*, 8(5).

Deshpande, M. and Karypis, G. (2000). *Selective Markov Models for Predicting Web-Page Accesses*. University of Minnesota,Army HPC Research Center Minneapolis.

Ingrid Zukerman, D. W. A. (2001). *Predictive statistical model for user modeling*. School of Computer Science and Software Engineering, Monash University, Clayton, Victoria.

James Pitkow, P. P. (1999). Mining longest repeating subsequence to predict world wide web surfing. *Second USENIX Symposium on Internet Technologies and Systems, Boulder, C0*.

Josep Domenech, Julio Sahuquillo, J. A. G. and Pont, A. (2012). A comparison of prediction algorithms for prefetching in the current web. *Journal of Web Engineering*, 11(1).

Konishi S., K. G. (2008). *Information Criteria and Statistical Modeling*. Springer, New York.

Kurian, H. (2008). *A MARKOV MODEL FOR WEB REQUEST PREDICTION*. Dr. Babasaheb Ambedkar Technological University.

Liu, B. (2006). *Web Data Mining: Exploring Hyperlinks, contents and Usage Data*. Springer, New York, 1 edition.

Martin Labsky, Vladimr Las, P. B. (2006). *Mining Click-stream Data With Statistical and Rule-based Methods*. Department of Information and Knowledge Engineering, University of Economics, Prague, Praha, Czech Republic.

Qiang Yang, H. H. Z. (2003). *Web-Log Mining for Predictive Web Caching*. IEEE Computer Society.

Zhong Su, Qiang Yang, Y. L. H.-J. Z. (2000). Whatnext: A prediction system for web requests using n-gram sequence models. *First International Conference on Web Information Systems and Engineering Conference*.

BIOGRAFI PENULIS

Penulis lahir di Jombang, putri ke-sembilan dari 10 bersaudara. Menempuh pendidikan dasar di SDN Jabon I Jombang . Melanjutkan pendidikan menengah pertama di SMPN I Jombang lulus pada tahun 1991.

Kemudian menyelesaikan pendidikan menengah atas di SMAN II Jombang jurusan Biologi (A2) lulus pada tahun 1994.

Tahun 1995 melanjutkan Pendidikan Tingkat Sarjana dengan Program Studi Teknik Informatika Fakultas Teknologi Informasi (FTIf) Institut Teknologi Sepuluh Nopember Surabaya lulus pada tahun 2002.

Tahun 2013 mendapatkan kesempatan untuk melanjutkan pendidikan tingkat magister di Jurusan Teknik Elektro ITS bidang keahlian Telematika konsentrasi Chief Information Officer (CIO) dengan beasiswa dari Badan Litbang SDM, Kementerian Komunikasi dan Informatika. Kontak Person melalui email elokswahyuni@gmail.com